Unit 10 Experiments

Contents

Introduction

Experimentation plays a critical role in the advancement of knowledge and the development of our society. Technological development in numerous areas, such as agriculture, electronics, manufacturing and medicine, depends to a greater or lesser extent on knowledge that has been collected from scientific experiments. This unit discusses the nature of experiments, and you'll learn statistical methods that are suited to the analysis of small sets of data. You will also actually conduct an experiment, giving you hands-on experience of collecting **empirical data** – that is, data collected through observation or experimentation.

Section 1 says more about the role of experiments in the world and the diversity of questions that can be addressed through experiments. Different types of experiment are also identified, focusing on hypothesis-testing experiments because these make the greatest use of statistics.

In Section 2, you are asked to set up an experiment on the growth of plants, using mustard (or cress) seeds. This should give you an idea of some of the problems that are involved in even the simplest of experiments.

Important: planning your schedule

The timing of the experiment in Section 2 is important, and you therefore need to plan your schedule accordingly.

- It will probably take you about one-and-a-half hours to set up the experiment, assuming that you have already collected the items in the list – posted on the website some weeks ago, and repeated in Subsection 2.2.
- You will have to spend a few minutes on your experiment daily for four or five days after setting it up, and then about one-and-a-half hours taking measurements from your plants.

Not all experiments are like this one, performed in a laboratory.

The analysis of the data from this experiment uses a new form of test, called the t-test, which is the major topic of Sections 3 and 4. The t-test has many similarities to the z-test. Like the z-test, its purpose is to test hypotheses about population means. Also, as for the z-test, there is a two-sample test for comparing two populations, and a one-sample test for testing hypotheses about a single population. We can also form confidence intervals related to t-tests in the

same way that we have with z-tests, as you will see in Section 5. The attraction of the t-test is that it can be used with samples of any size, including small sizes – unlike the z-test, which we use only with samples of size 25 or more.

With both z-tests and t-tests, the alternative hypothesis is usually 'two-sided' – the null hypothesis specifies the value of the population mean, and we reject this hypothesis in favour of the alternative hypothesis if the sample mean differs substantially from that hypothesised value. Occasionally, though, a 'one-sided' alternative hypothesis is appropriate, where we would only reject the hypothesised value if the difference were in a particular direction. (Perhaps we would not reject the null hypothesis if the sample mean turned out to be much bigger than the hypothesised value, but only if it were much smaller.) In Section 6 we look at one-sided alternative hypotheses.

Section 7 directs you to the Computer Book, where you will learn to use Minitab to perform t-tests and to calculate the associated confidence intervals.

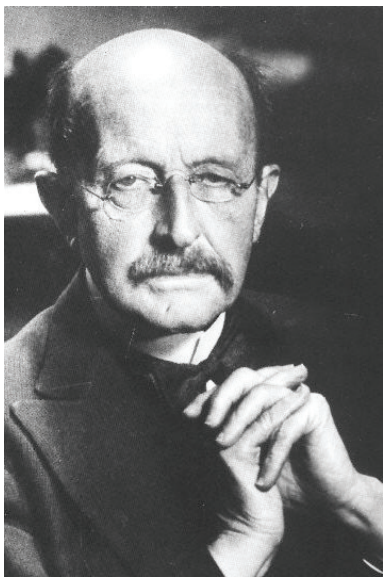
1 Scientific experiments

In this section we first describe those forms of enquiry for which we use the term *experiment*. We then describe three different kinds of experiment that are common in scientific enquiry. One of these is looked at in more detail – the one which most commonly involves statistical analysis.

1.1 What are experiments?

> 'An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer.'
>
> Max Planck (1858–1947)

The term **experiment** means a variety of things to a variety of people. To many it conjures up a vision of a white-coated individual surrounded by dials, flashing lights and vaporous fluids bubbling sullenly in mysteriously coiled vessels. To others an experiment involves no more than adding a new ingredient to a tried and trusted pie recipe to see if the taste is improved, or planting the bulbs earlier than in previous years to see if they do better. These uses are all perfectly proper: it is not *what you do* that qualifies an activity as an experiment; it is *the way that you do it*.



Max Planck

In other words, the area in which you carry out an experiment might be nuclear physics or it might be cookery: neither lies outside the province of experimentation. But if other people are going to accept that you have carried out an experiment, then you must use certain methods and procedures.

One fundamental feature of any experiment is that it should *stand up to scrutiny*. Sometimes the only person interested in the result of an experiment is the person conducting it. For example, a golfer may have a set of lessons or experiment with his choice of golf clubs in order to improve his score. Quite possibly the outcome is only of interest to the golfer. However, for this to be an experiment, the golfer must gather information that is factual and would enable others to critically evaluate whether his golf has improved. To be of any value to others, though, the results must also generalise – we would want to know whether golf lessons are typically beneficial. To meet that aim, an experiment must be **repeatable**. If person *A* carries out an experiment, then he or she should be able to explain everything that took place during the experiment in such a way that another person (*B*) could, if necessary, go through exactly the

Company or industrial confidentiality means that experiments and their results may not be in the public domain, but the company or industry will want that information.

same procedure. The results of this experiment should, again, be suitable for scrutiny so that *B*'s results can be compared with *A*'s.

The importance of recording the detail of an experiment is illustrated in procedures for developing a new drug, as you will learn in Unit 11. If a drug company carries out a clinical trial on a new drug, then they must be able to describe every part of their procedure to the outside world, including: how the experiment was designed; how many subjects were involved; how the drug was administered and in what quantities; how its effects were measured; etc. In fact, the European Commission *requires* such information before they grant a product licence. In just the same way, a cook who experiments by altering an ingredient in a recipe should be able to explain every part of the revised recipe so that other people could repeat the revised procedure exactly.

Another fundamental feature of an experiment is that *it sets out to answer a specific question or set of questions*. Does this drug alter blood pressure? Does this ingredient improve the taste of the pie? Does planting the bulbs earlier produce better flowers? Notice that each of these questions is framed in such a way as to demand that something be measured if the question is to be answered: namely, blood pressure, taste or flower quality. The questions may sometimes seem vaguer. For example: 'What happens in the long term from taking statins daily?' However, decisions must be made on which measurements are taken and what information is recorded, and these choices sharpen a vague question. Thus, blood pressure might be recorded in an experiment to examine the affects of statins. Then one question is: 'Do statins affect blood pressure?' Or weight might be monitored, or incidence of strokes or diabetes. In each case, examining this information in the context of the experiment implies a question.

Activity 1 Pie-tasting

A professional chef wants to discover if the addition of an extra ingredient improves the taste of a pie. Describe a suitable experiment to find this out.

This activity shows that it is possible to carry out scientific experiments on all sorts of things, not just on formally recognised scientific subjects. To illustrate the difference between the scientific approach to a question and other forms of inquiry, it is instructive to consider how a scientific experiment might be carried out on such an unlikely subject as poetry appreciation. Take a poem such as John Keats's ode *To Autumn*. (A copy of the poem is available on the M140 website.)

A literary critic might ask the question: 'Is *To Autumn* a great poem?'. Framed in this way the question is not amenable to scientific experiment. People might vary in their opinion of the poem, and they would probably produce more or less convincing evidence to support their viewpoint. One might point to the images that the poet uses and argue that they successfully convey the mood of autumn, another might complain that the overall effect is too languid for his taste, etc. Although people might agree as to what are good and bad criteria for judging the greatness of a poem, and although informed opinion might generally agree as to the greatness of this particular poem, there is no sense in which assessing the greatness of this poem is a scientific experiment. However, it is possible to ask questions about the poem which are open to experimental investigation.

Activity 2 Two or three verses?

The poem *To Autumn* has three verses. Someone might argue that the third verse is inferior to the others, and that without it the poem is greatly improved. How might you test this experimentally?

Of course nobody would imagine that this particular experiment is of any literary value. Our purpose is simply to show that experimental investigation need not be limited to the physical world.

1.2 Different kinds of experiment

Good experiments share two important features:

- They are recorded in detail so that they can be critically evaluated and repeated.
- They produce measurements or observations designed to answer specific questions.

There are different kinds of experiment, addressing different kinds of question. This is not a case of, for example, some experiments being about physics and others about chemistry; it is a case of the questions being different in their purpose.

Some experiments answer questions of the form:

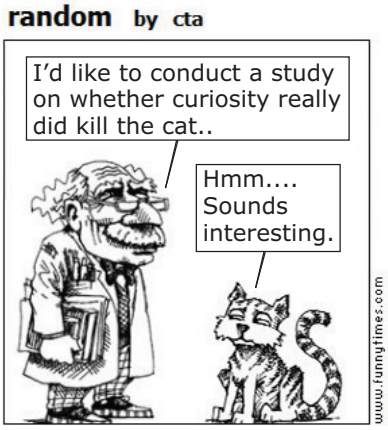
What happens if I do this?

Some of the experiments that you will meet in Unit 11 answer questions of this sort. For example, the following question might arise in the early stages of drug testing:

What happens if I give this person this new drug?

Such experiments are **exploratory** in nature, in the same way that toddlers' investigations of their surroundings are exploratory.

Francis Bacon, a sixteenth-century scientist and philosopher, urged his contemporaries to carry out experiments of this type, and for this reason you may find such exploratory experiments referred to as **Baconian**.



Sir Francis Bacon

Sir Francis Bacon (1561–1626) was a Renaissance thinker and an English statesman. He was a member of parliament at the age of 23, went on to be Attorney General and Lord Chancellor, was knighted, made a baron, and later made a viscount. However, his most enduring legacy is his contribution to scientific method. Bacon established and popularised *inductive* methodologies for scientific inquiry. By 'induction', Bacon meant the ability to gradually generalise a finding based on accumulating information. He argued that, to learn about nature, data should be gathered through organised experiments that provide tangible information and increase knowledge.

Francis Bacon (1561–1626)

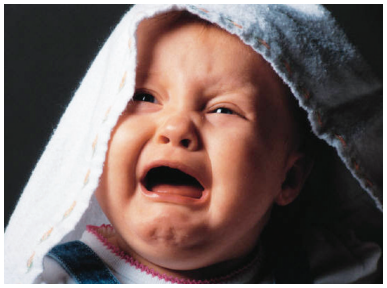
A second kind of experiment has, as its primary purpose, not exploration but the **measurement** of a particular attribute. Such experiments are designed to answer questions such as:

- What is the velocity of light?
- How heavy is the Earth?
- How old is this rock?
- What is the population of the UK?

This kind of experiment is a very important part of scientific investigation, but we shall not discuss such experiments at length in this unit.

A third kind of experiment, however, is both important and relevant to this module. This is the kind that tests a specific *hypothesis*. Perhaps the best way to explain this kind of experiment is to give some examples.

Example 1 Crying baby



Hypothesis: The baby is crying because it is cold.

Prediction based on hypothesis: If the baby is made warmer, then its crying will stop.

Test: Make the baby warmer, but make sure that you do not change anything else in its environment.

Possible results and conclusions: The baby continues crying; therefore the hypothesis is wrong. The baby stops crying; therefore the hypothesis is supported. (Since you cannot discount the possibility that the baby would have stopped crying anyway, you cannot say that the hypothesis is correct.)

Example 2 Broken TV

Hypothesis: The television is not working because the fuse in the plug is broken.

Prediction based on hypothesis: If the fuse is replaced with one that does work, then the television will work again.

Test: Replace the fuse but do not alter anything else.

Possible results and conclusions: The television still does not work; therefore the hypothesis is wrong. The television works; therefore the hypothesis is supported.

Example 3 Are microbes to blame?

Hypothesis: Food putrefies if left for too long because of the action of microbes (small organisms) that are present in the air and that come into contact with it.

Prediction based on hypothesis: If the microbes are prevented from acting on the food, then it will not putrefy.

Test: Prevent the microbes from acting by killing them or otherwise preventing them from acting (for example, by deep-freezing).

Possible results and conclusions: The food putrefies; therefore the hypothesis is wrong. The food does not putrefy; therefore the hypothesis is supported.

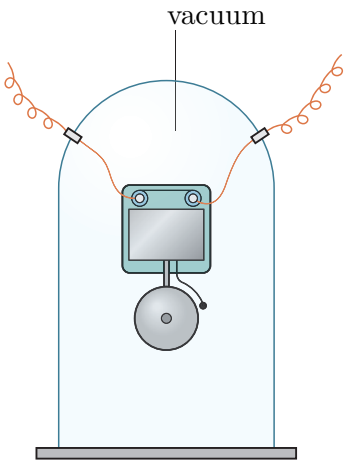
Example 4 Transmission of sound

Hypothesis: Sound is transmitted through air because sound is a form of mechanical vibration and air contains particles (molecules) that can jostle each other and so pass the vibration from one particle to the next.

Prediction based on hypothesis: If all the air is removed from a container so that a vacuum remains, then it should not be possible to transmit sound through that vacuum. For example, it should not be possible to hear an electric bell ringing inside a container from which the air has been removed.

Test: Remove air from a container and discover whether an electric bell inside it becomes inaudible.

Possible results and conclusions: The bell can be heard through the vacuum; therefore the hypothesis is wrong. The bell cannot be heard; therefore the hypothesis is supported.



Bell in a vacuum

The first two of these examples are homely examples of **hypothesis-testing experiments** of a kind that people routinely carry out in their daily lives. The latter two examples are well-known, formal, scientific experiments. The experiment in Example 3 was first carried out by Louis Pasteur (1822–1895), who successfully prevented microbes from attacking food by filtering the air in which the food was kept, and by taking food to the tops of mountains, where the combination of a low temperature and relatively microbe-free air prevented the food from decaying. Example 4 dates back to Athanasius Kircher (1601/1602–1680), who wrote in 1650 about experiments with bells in a vacuum – now a staple of physics lecture demonstrations.

Louis Pasteur (1822–1895)

All four of the examples have the following important features in common.

- *Hypothesis:* In each, there is a specific hypothesis about the cause of a phenomenon. A hypothesis-testing experiment tries to explain how something works.
- *Prediction based on hypothesis:* The experimenter makes predictions that flow directly from the hypothesis – if the hypothesis is true, then certain things must follow. For example, if it is true that microbes are the sole cause of food decay, then it automatically follows that food should not decay if microbes are absent.
- *Test:* Each experiment consists of testing whether the things that the hypothesis predicts actually happen.
- *Possible results and conclusions:* If the result of the experiment is not what the hypothesis predicted, then the experimenter has to accept that the hypothesis is wrong. For example, if food were to continue to decay, even when it was absolutely certain that no microbes were present, then the hypothesis that microbes are the only cause of food decay would have to be rejected.

It is important to note that even if the prediction that the hypothesis makes turns out to be correct, then it may still be wrong to assume that the hypothesis itself is perfectly correct.

Consider, for example, the hypothesis that malaria is caused by a mosquito (more specifically, by a particular type of mosquito of the *Anopheles* genus). A prediction which follows from this hypothesis is that malaria should cease to occur in a district from which *Anopheles* mosquitoes have been eradicated. In fact, this is what normally happens in practice, so it would seem reasonable to

conclude that the hypothesis is correct. However, although it is correct in one sense, it is not in another. In a district from which the *Anopheles* mosquito has been eradicated, some people may still continue to contract malaria, apparently spontaneously, long after the mosquitoes have gone. If you were ignorant of these cases of malaria, then it would be legitimate to believe in the original hypothesis that *Anopheles* mosquitoes cause malaria. As soon as these few instances come to light, the original hypothesis must be discarded and a new one sought.

In the case of malaria, the truth is that the mosquito is only a carrier for a parasite, called *Plasmodium*, which is the direct cause of malaria. *Plasmodium* can survive for long periods in the human body without giving rise to malaria. From time to time, however, it invades the blood, and malaria then develops.

To summarise, if the predictions that follow from a hypothesis turn out to be incorrect, then the hypothesis has to be abandoned or at least modified. If the predictions turn out to be correct, then the hypothesis is supported in the sense that it can be provisionally accepted that the hypothesis is correct. There is always the possibility that, one day, somebody may find that under certain conditions the predictions are incorrect. When that happens, the hypothesis has to be replaced by a new one.

1.3 Experiments and statistics

Before going any further, it is important to consider how this description of scientific hypothesis-testing experiments fits in with the ideas of statistical hypothesis-testing in Units 6 and 7. From the statistician's point of view, the examples of hypothesis-testing experiments in Examples 1 to 4 (Subsection 1.2) are framed in rather unconventional terms, because each concentrates on a hypothesis of the form *something affects something* and tests predictions flowing from that hypothesis. Nevertheless, it is perfectly possible to formulate each of the above experiments in statistical terms with null and alternative hypotheses, and it is essential to do so if you want to apply statistical hypothesis tests to data arising from such experiments. The null hypothesis, as its name implies, very often postulates the absence of a given effect or relationship.

When we carry out a statistical hypothesis test, we base our calculations on the assumption that the null hypothesis is correct. We ask:

> *What is the probability of obtaining a result at least as extreme as that which we have obtained, if we assume that the null hypothesis is true?*

If this probability is too low, then we reject the null hypothesis in favour of the alternative. (See Section 4 of Unit 6.)

This process is summarised by the flow chart given in Figure 1.

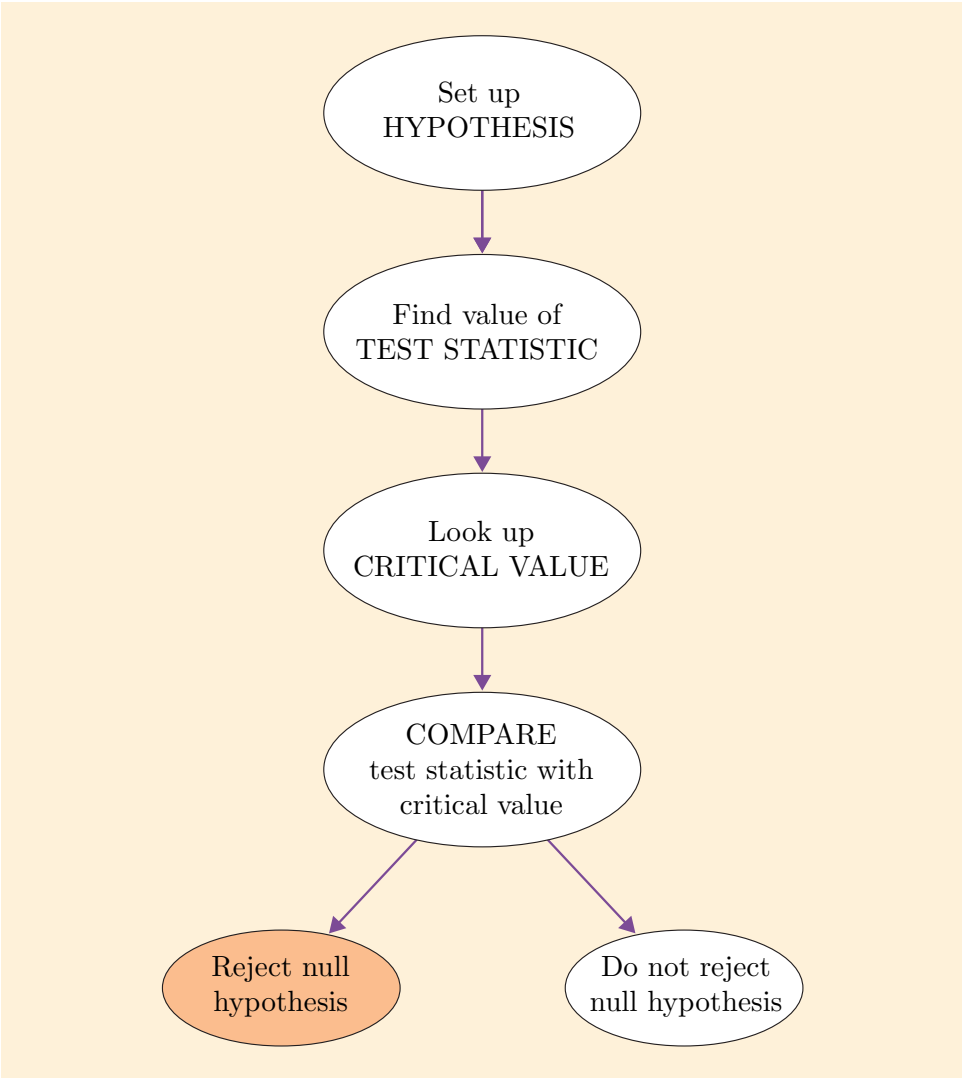


Figure 1 Steps in a hypothesis test

We shall now devise null and alternative hypotheses for the first two of the four experiments (Examples 1 and 2, Subsection 1.2). In each case, we shall assume that the tests are the same as those described earlier and we will state what conclusions we draw from the different results that could be obtained. In these examples, as often happens, the null hypothesis in the *statistical* hypothesis test is the converse (opposite) of the hypothesis in the scientific experiment. So then, in the test, the amount of evidence against the null hypothesis is assessed (see Subsection 1.3 of Unit 7). As a result, the reformulation is often not logically equivalent to the original experiment.

Example 5 Example 1 revisited

Null hypothesis: Cold has no effect on the baby's crying.

Alternative hypothesis: Cold makes the baby cry.

Test: Make the baby's surroundings warmer and set up suitable controls.

Possible results and conclusions: The baby's crying stops; therefore reject the null hypothesis. The baby's crying continues; therefore the null hypothesis is supported.

Example 6 Example 2 revisited

Null hypothesis: The present condition of the fuse is not responsible for the television's refusal to work.

Alternative hypothesis: The television is not working because the fuse is broken.

Test: Replace the fuse but do not alter anything else.

Possible results and conclusions: The television works; therefore reject the null hypothesis. The television still does not work; therefore the null hypothesis is supported.

Activity 3 Forming null and alternative hypotheses

Express the following experiments as statistical hypothesis tests.

(a) Example 3 experiment: Are microbes to blame?

(b) Example 4 experiment: Transmission of sound.

One final point needs to be made about the relationship between hypothesis-testing experiments and statistical hypothesis tests. A great deal of scientific experimentation consists of testing specific hypotheses of the sort just described. If the experiments require statistical analysis, then the experimenter should use statistical hypothesis tests. This will be the case if the experiments involve things that are intrinsically variable, such as people, plants or animals. Sometimes, however, a scientist may be interested not so much in testing whether or not a given treatment has an effect (perhaps somebody has already done an experiment which shows that it does) but rather in investigating how big that effect is. An agriculturalist might want to know, for example, by how much a new fertiliser increases the yield of a crop. In such instances the scientist conducting the research project would use statistical estimation procedures (for example, finding a confidence interval) rather than carrying out a hypothesis test.

Activity 4 Identifying scientific experiments

State whether each of the following is a scientific experiment.

(a) Measuring the distance between the Earth and the Sun.

(b) Leaving work an hour later to see if it makes much difference to your travel time to get home.

(c) Reading a tea-taster's report on a brand of tea in order to decide whether to buy it or not.

(d) Investigating whether obesity is caused by overeating.

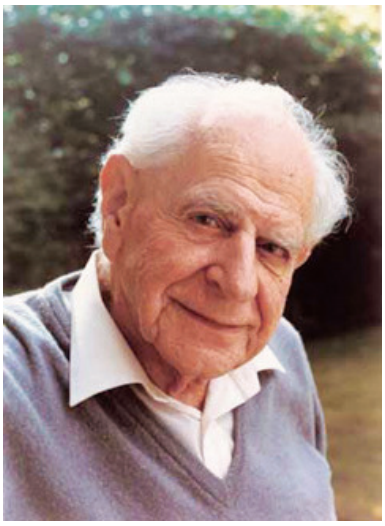
Activity 5 Types of experiment

For each of the experiments in Activity 4 that you identified as scientific, state which of the three kinds it is: exploratory, measurement or hypothesis-testing.

Activity 6 Forming statistical hypotheses

For those scientific experiments in Activity 5 which you identified could be hypothesis-testing experiments, formulate them as statistical hypothesis tests.

Hypothesis-testing experiments involve testing deductions that can be made from a hypothesis. For this reason, scientific experimentation is sometimes said to proceed by the **hypothetico-deductive method**. This description of scientific method is frequently associated with the philosopher Karl Popper (1902–1994). There has been considerable debate about whether science does proceed by this method under all circumstances, but this debate lies outside the scope of this module. Many scientists feel that carrying out an experiment is rather like riding a bicycle: if you think too hard about what you are doing, you fall off! Accordingly, it is probably most profitable to leave such discussions of the nature of scientific experimentation and to get down to the practicalities of carrying out an experiment.



Karl Popper (1902–1994)

Kinds of experiment

In summary, there are at least three kinds of experiment that can be recognised: they are distinguished by the kind of questions they attempt to answer.

- Exploratory (Baconian) experiments
- Measurement experiments
- Hypothesis-testing (hypothetico-deductive) experiments

Exercises on Section 1

Exercise 1 Scientific experiment?

State whether each of the following is a scientific experiment.

(a) Driving a car with all the windows open to see whether petrol consumption is affected.

(b) Viewing a television advert for car insurance in order to decide whether to purchase the insurance or not.

Exercise 2 What form of experiment?

(a) For the scientific experiment identified in Exercise 1, state which of the three kinds it is: exploratory, measurement or hypothesis-testing.

(b) Formulate this scientific experiment as a statistical hypothesis test.

2 Carrying out your own experiment

Now that you have read about some of the basic principles of experimentation, it is worth trying to put these principles into practice by setting up and carrying out your own experiment – as detailed in this section.

You should read the whole of Section 2 before you start the experiment.

The experiment is concerned with the growth of mustard (or cress) seedlings. Mustard is better, because it grows faster, but cress will do perfectly well. You should set up the experiment and then leave the seedlings to grow for four or five days, checking them from time to time and pruning some of them after two or three days. Provided that you have already collected all the necessary items (as detailed in Subsection 2.2), it should not take you more than about one-and-a-half hours to set up the experiment. You will then need to take measurements as part of the experiment, which is likely to take another one-and-a-half hours. The details of these activities are in Subsections 2.3 to 2.6.

After you have taken the measurements and read about t-tests in Section 3, you will be ready to analyse your results statistically.

2.1 Purpose of the experiment

1. Clarify

The purpose of the experiment is to discover whether light affects the growth of plant roots. Most people are familiar with the fact that light is important to plants, and realise that it affects the growth of the stems and leaves. If you keep a potted plant on a window sill, for example, and never turn it round, then it is likely to grow quite noticeably towards the light. But what effect, if any, does light have on the growth of the roots? Normally, of course, the roots are underground and so are not exposed to the light, but it is quite conceivable that the roots could become exposed as they grew – for example, if the plant was growing on an irregular surface or if rain washed some of the soil away. Under such circumstances the plant would probably not benefit from the root continuing to grow out into the open. Roots give plants mechanical support, and they absorb water and nutrients from the surrounding soil. They can do none of these things if they are growing in the open air. It seems a plausible hypothesis, therefore, that light suppresses root growth so that the roots will tend not to grow in the open air. (It is also possible to argue plausibly that the roots will tend to grow longer in the open air.)

The experiment which you are asked to carry out is on mustard seedlings. Thus the precise question to be answered is:

Does light affect the root growth of mustard seedlings?

2. Collect

The principle of the experiment is simple. You are asked to grow two groups of mustard seedlings, one entirely in the dark and the other entirely in the light, but otherwise in as near identical conditions as possible. After some time has elapsed you should measure the lengths of the roots of the seedlings and compare the two groups. As in any experiment, it is essential to control various factors.

Activity 7 Controlling factors that might undermine the experiment

List some factors which need to be controlled, and suggest how such control might in each case be achieved.

This experiment should provide a reasonable amount of data. Before going on to analyse the data from your experiment, you will need to consider the hypothesis being tested (as in Section 1).

Activity 8 Hypotheses and potential results

State the null hypothesis that is being tested by this experiment. State the prediction that follows if the null hypothesis is correct. State the conclusions that you can draw from the different results which you might get.

4. Interpret

If you were to discover, however, that the difference appears only in seedlings whose stems have not been cut, then you should suspect that light does not directly affect root growth, but does so indirectly through its effect on the leaves and stem.

2.2 Items needed for the experiment

Figure 2 The items needed for the experiment

As shown in Figure 2, you will need the following:

- Two identical plastic containers such as small flower pots or empty cartons, e.g. of margarine or yoghurt. These need to be at least 6 cm (2.5 inches) in diameter at the open end or, if rectangular, need to have sides at least 5 cm (2 inches) long.
- Two large containers (ice cream tubs, sandwich boxes, small buckets or large bowls) that will hold at least half a litre (or 1 pint) of water without leaking, and which can each hold one of the plastic containers mentioned above (see Figure 4, Subsection 2.3). These larger containers will need to stand side by side in a well-illuminated position, e.g. on a window ledge.
- One small packet of mustard seeds (*Sinapis alba, Brassica alba* or *Brassica hirta*) – various companies supply these. (If these are difficult to obtain, then use cress seeds (*Lepidium sativum*).)
- One piece of aluminium kitchen foil about 30 cm $\times$ 30 cm (12 inches $\times$ 12 inches).
- One piece of clear plastic (e.g. from a plastic bag) about 30 cm $\times$ 30 cm (12 inches $\times$ 12 inches) or a clear plastic bag.

- Some pieces of (superior quality) toilet tissue, or kitchen roll.
- Two sheets of high quality (printer or writing) paper, preferably coloured so that the seedlings show up against the paper.
- Four elastic bands to go round the tops of the flower pots or margarine or yoghurt cartons.
- A jug or bottle for pouring about half a litre (or a pint) of water into the large containers.
- A pair of dividers or a piece of plasticine (or a similar material) plus two pins.
- A ruler measuring millimetres.
- Two teaspoons.
- A pair of fairly small, sharp-pointed scissors.
- A magnifying glass would be helpful but is not essential.

2.3 Setting up the experiment

Here are the instructions for setting up the experiment.

1. Empty all the seeds (i.e. at least 40) into a bowl containing cold tap water (see Figure 3), and leave them to soak for an hour. (Seeds remain dormant while they are dry; soaking them ensures that all the seeds start germinating at the same moment and that they all get off to a good start.)

Figure 3 Soaking the seeds

2. Make sure that the two large and the two small containers which you are going to use are thoroughly clean, and thoroughly rinsed free of detergent, soap or any other contaminant.
3. Put half a litre (or a pint) of tap water into each of the large containers.
4. Take the two small containers, and if they do not already have holes in the bottom, cut one or two so that the water can easily seep up through the holes (see Figure 4).

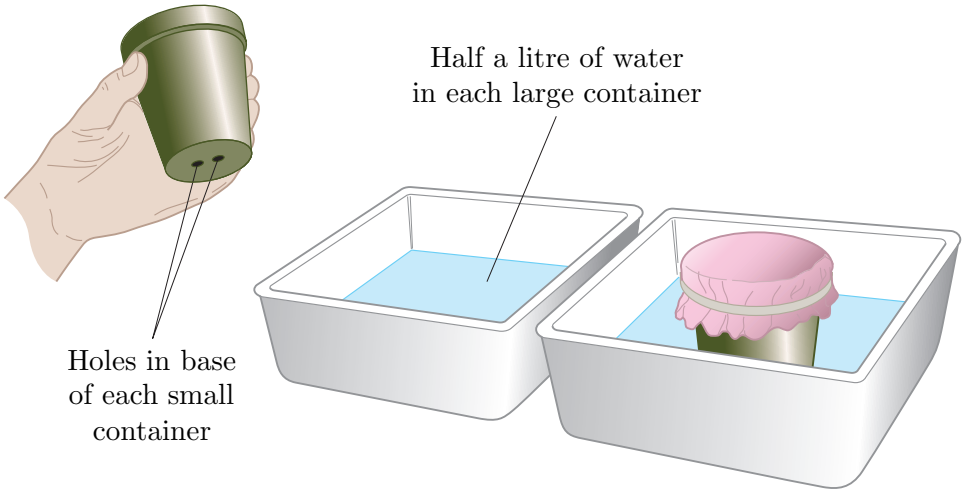


Figure 4 Flower pots and plastic containers

5. Crumple several lengths of toilet tissue or kitchen roll, and press them firmly into each small container. Continue until both containers are just overfull. The paper should be firm but not jammed solid.
6. Fold several lengths of tissue paper to make a smooth platform, then lay a sheet of high quality (printer or writing) paper on top. Ensure that this platform touches the crumpled tissue paper. Fix this platform of paper over the top of one of the small containers with an elastic band (see Figure 5). Repeat for the other small container.
7. Stand each small container in one of the large containers (see Figure 4).
8. By the time that the seeds have been soaked for one hour (see Instruction 1), the paper in each small container should be thoroughly damp. If it is not, then gently spoon over the paper surface some of the water from the big container in which the small container stands.
9. Reject any seeds that seem unusual – for example, those that are a different colour from the rest and those that float rather than sink in the water.
10. Mix the seeds well and then allocate them alternately to two groups until 20 have been allocated to each.

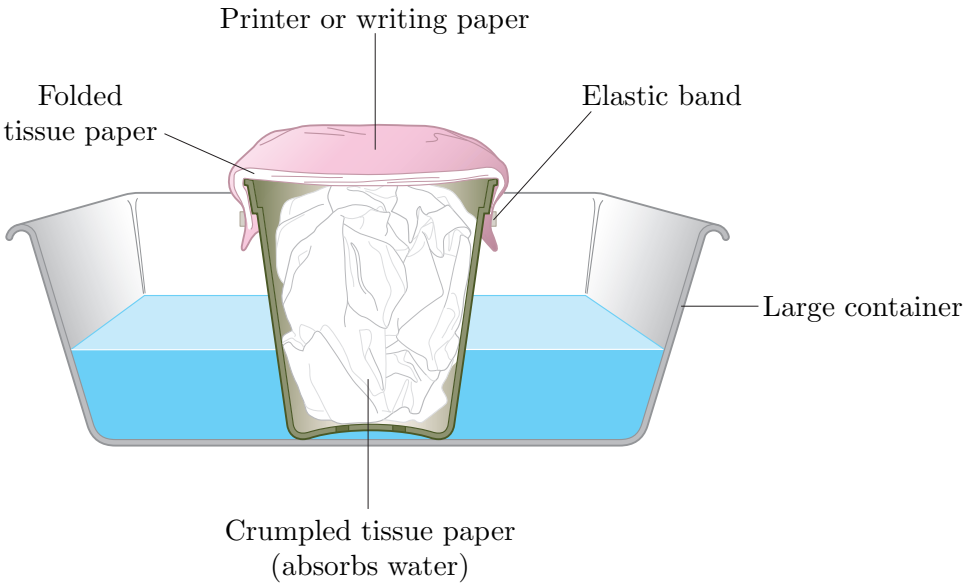


Figure 5 Setting up the platform

11. Arrange the two groups of 20 seeds as follows, one group on the writing

paper surface over each small container. A teaspoon may help you to manoeuvre the seeds into position. Arrange each group of seeds in four rows of five, each row being at least 1 cm (half an inch) away from the next. The further apart you can put them, provided that they are spaced regularly, the better (see Figure 6).

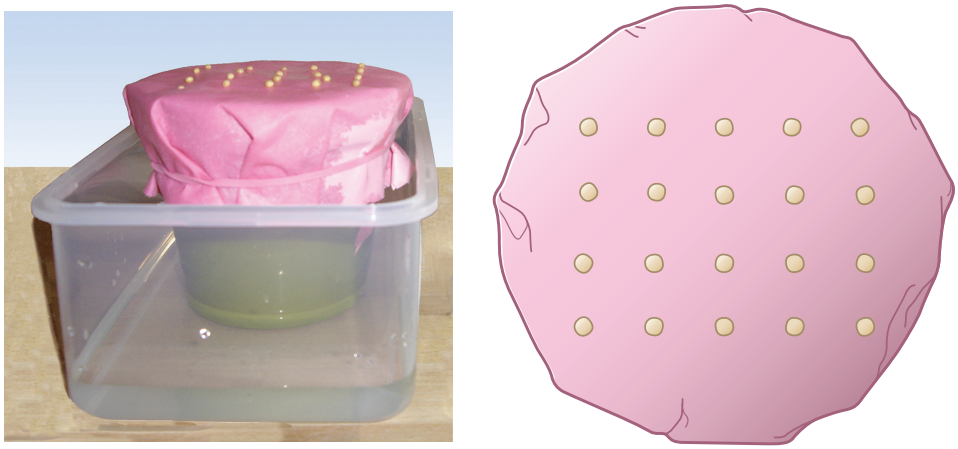


Figure 6 Spacing of seeds

12. Toss a coin to select, at random, the pot whose seeds will grow in the dark; then make a hood out of the aluminium foil and wrap it over the top of the chosen pot, leaving a space of at least 5 cm (2 inches) between the seeds and the top of the foil (see Figure 7). Secure the foil with an elastic band. Repeat the hood-making procedure, using the piece of clear plastic or plastic bag, for the second pot.

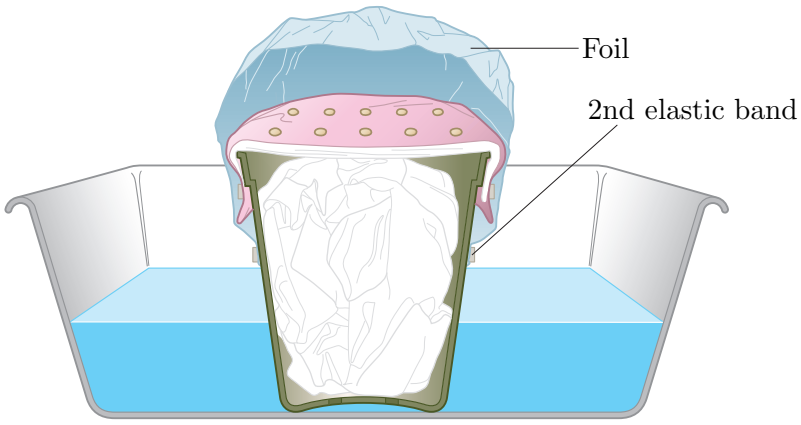


Figure 7 Putting a hood on the pots

13. Place the two sets of containers side by side in a well-lit spot safe from disturbance by children, pets or curious passers-by. Try to ensure that the temperatures of the two groups are the same.

This completes the setting-up procedure.

2.4 Maintaining the experiment

You also have to undertake a few tasks during the days that follow. Exactly how fast the seedlings develop will depend on the temperature, but you can expect to cut the stems off some of the seedlings (see Subsection 2.5) after about two to five days, and to measure the root lengths after three to seven days. The maintenance tasks are as follows.

1. To control for any difference in the temperature of the two groups of seedlings, swap the positions of the two large containers each day.
2. Make sure that there is still plenty of water in the large containers. If they begin to dry out, add another half a litre (or a pint) of water to each of the large containers.
3. Check the seeds which you can see (i.e. those growing in the light) once a day to see how rapidly the stems and roots are developing.

2.5 Cutting the stems

When most of the seedlings growing in the light have grown root hairs and have stems which are a little over 1 cm (about half an inch) long, cut the stems off 10 seedlings in each pot (i.e. 20 seedlings in all). Any seedling which has not grown sufficiently for its stem to be cut should be left uncut.

Figures 8 and 9 show you where to cut the seedlings.

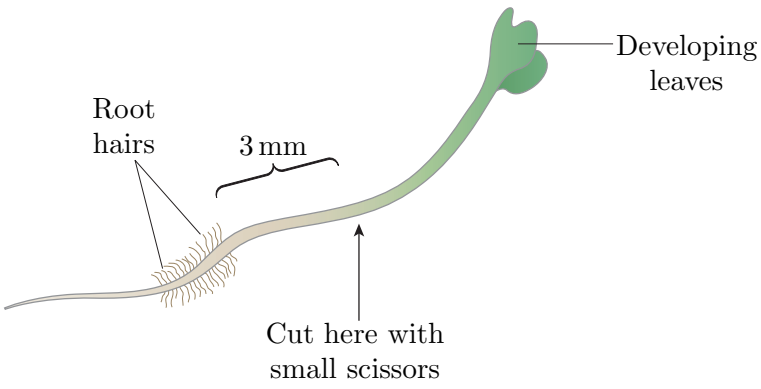


Figure 8 Where to cut seedlings

You will see on most of the roots a white, fluffy area which is due to the presence of a multitude of very fine hairs, called root hairs. It is primarily through these hairs that the root absorbs water and nutrients. Use the top of the root, where the root hairs stop, as a reference point, and cut across the stem about 3 mm ($\frac{1}{8}$ inch) above this point. Cuts should be made without moving the seedlings, using small, sharp-pointed scissors. Any seedling which has not produced any root hairs should be left uncut.

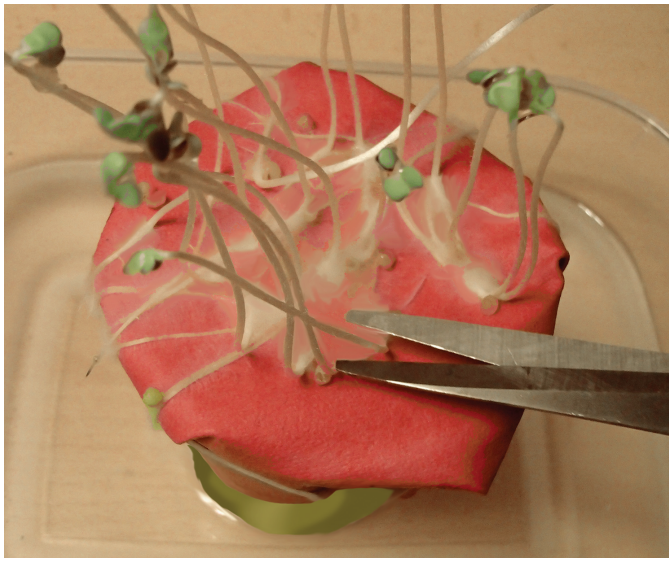


Figure 9 Cutting a stem

When you have finished cutting the stems, replace the aluminium foil and clear plastic/plastic bag on the pots they were previously on, taking care not to squash the seedlings, and return them to their positions (in the large containers) by the window. If the cut stems start to grow rapidly, then repeat the cutting process a few days later.

2.6 Completing the experiment

Leave the pots until the roots of most of the seedlings in the light have grown to at least 1 cm (about half an inch) long. Some seeds may not germinate at all, and these should be ignored at this stage.

To measure the root lengths, you will need either a pair of dividers, a pair of compasses, or a piece of plasticine and two pins, or needles, assembled as shown in Figure 10. This is the trickiest part of the experiment. Start with one pot of seedlings and, working along each row in turn, carefully lift each seedling off the paper, taking great care to ensure that the tip of the root is not left behind. Lay the plant down on a flat, preferably dark-coloured, surface and straighten the root as far as possible. Measure from the tip of the root to the position near the top of the root where the root hairs stop (see Figure 10). You should do this by putting one point of the dividers etc. against the root tip and then moving the other until it lies opposite the position where the root hairs stop. Then measure the distance between the two points of the dividers etc. with the ruler (see Figure 11). Measure it in millimetres (to the nearest whole millimetre). You may find it difficult to measure the root lengths in this way; if you do, then try measuring the roots directly with the ruler instead.

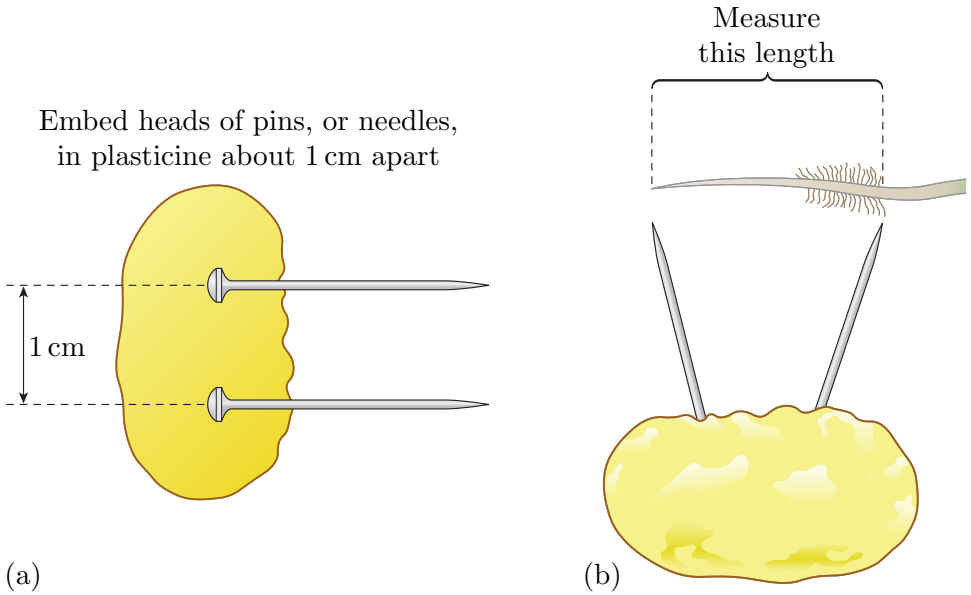


Figure 10 Measuring tool and measuring root lengths

Write the measurement down in the appropriate place in Figure 12. For each seedling, indicate very clearly whether or not you have previously cut its stem (e.g. underline the measurements of seedlings whose stems had been cut). You will also need to indicate any seeds that failed to germinate by putting a cross in the appropriate box. Measure all the seedlings in one pot in this way and then repeat the procedure for the second pot.

2. Collect

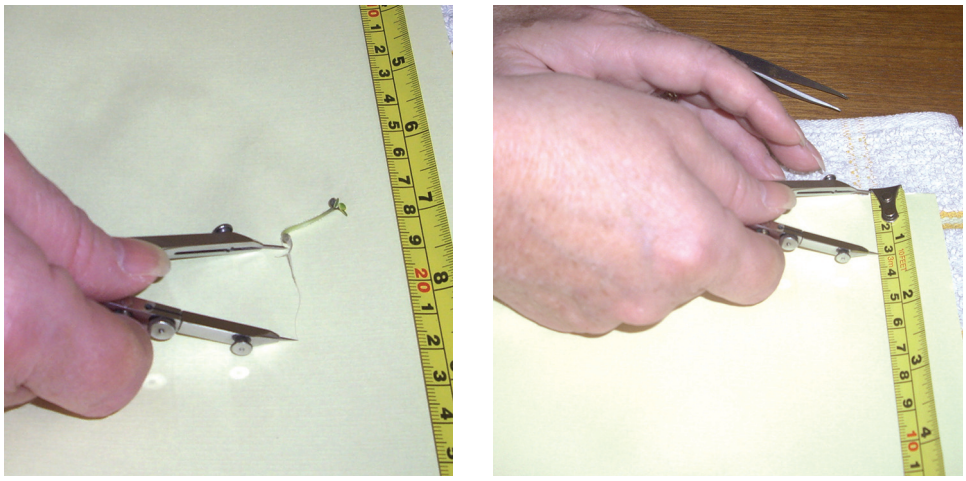


Figure 11 Measuring a root length

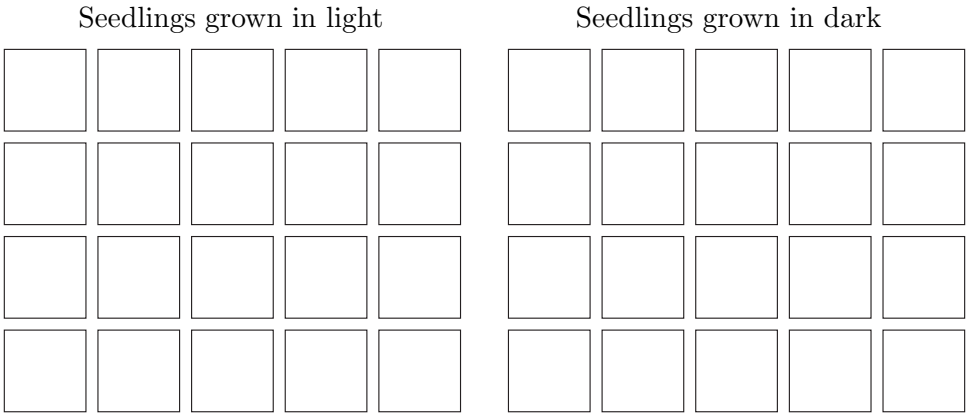


Figure 12 Grids for recording your results

2.7 Variability, error and clarifying the question

Whatever differences might exist between your two groups of seedlings, it is likely that they will not be very large. If they are large, then there would be no need to use a hypothesis test to analyse the data. One of the reasons why any differences that might exist may not be obvious, despite all the controls which you have used, is that the seedlings themselves are variable. You may try very hard to ensure that all the seedlings grow in the same conditions, but nevertheless there will be some small variations that affect the plants. More importantly, even if they are grown under identical conditions, different plants will still grow differently. We have also explained that you might find some difficulty in measuring the lengths of the roots accurately. In particular, because you may not be able to straighten the roots perfectly, it is possible that you will consistently underestimate their length. This kind of error, which is consistent in its direction and approximately constant in its magnitude, is called **systematic error**.

Most scientific experiments are subject to both variability and systematic errors. The task of the scientist (i.e. you) in carrying out an experiment is to discover, despite the presence of variability, bias and systematic errors, whether genuine differences exist between the two groups. One major source of variability, which you are very likely to experience, is that some of your seeds may not germinate at all. Later in this unit, when you calculate the means of the root lengths for your two groups of seedlings, should these non-starters be included?

There is no hard-and-fast rule as to how to deal with a problem like this, but it would seem sensible in this experiment to exclude them. Thus the aim of the experiment is to answer the following, even more precise, question.

The question to be addressed

Does light affect the root growth of those mustard seedlings that germinate?

1. Clarify

Now you should set up the experiment as described in Subsection 2.3. While you are maintaining it as described in Subsection 2.4, you should work through Section 3 to learn how to analyse the data that you will collect.

3 The t-test for two unrelated samples

The last section gave a procedure for setting up and running a scientific experiment to investigate the growth of mustard seedlings. The experiment yields data that are related to the question *Does light affect the root growth of those mustard seedlings which germinate?* You now need to decide how to analyse these data.

In this section we shall first consider how to do this in the context of the hypothesis tests that were introduced in Units 6 to 8. Then we shall introduce a further test which is particularly useful for the kind of data that we will have.

3.1 Which test?

In Units 6 to 8 we introduced some hypothesis tests: the sign test, the χ^2 test and the z-test. We shall now apply the principles of hypothesis testing developed there to the data that you will obtain from your mustard seedlings.

Activity 9 Appropriate hypothesis test?

Of the tests that you have already met in the module, is there one that would be appropriate for analysing your seedling data?

Is there a test that can be used on small samples of data?

The answer is a qualified yes: there is such a test, but it can be applied only to data from populations satisfying certain distributional conditions. These conditions will be described in Subsection 3.3. They include a requirement that the two populations must have the same variance. The other conditions are similar to ones you have met earlier in M140. (In Unit 12, you will meet a version of the test which does not require the population variances to be equal.)

Amongst the tests introduced in Units 6, 7 and 8, the one that comes closest to what is needed is the two-sample z-test. This test can be used to compare two unrelated samples of measurements, as with your seedlings data, but it can only be used for large samples of data, which rules it out. Nevertheless, it is worth looking again at the rationale behind the z-test to see why it requires large samples. We shall then give the test that you should use for your data. The test has close similarities to the z-test.

3.2 The z-test reconsidered

Suppose that you have unrelated samples of data from two populations, whose population means are μ_A and μ_B, and that you want to use the two-sample z-test to test the following null hypothesis H_0 against the alternative hypothesis H_1:

$$H_0 : \mu_A - \mu_B = 0$$
$$H_1 : \mu_A - \mu_B \neq 0.$$

Another type of z-test

The rationale of the z-test is as follows.

1. If the null hypothesis were true, then the means, μ_A and μ_B, of the two populations would be identical, so on average you would expect the means from two samples, one taken from each population, to be the same. In other words, on average you would expect the difference between the means of the two samples to be 0. If the means of the two samples are denoted by $\overline{x}_A$ and $\overline{x}_B$, respectively, then on average you would expect that $\overline{x}_A - \overline{x}_B = 0$.
2. If the null hypothesis were true and you took repeated pairs of samples from two populations like this, then you would find that sometimes $\overline{x}_A$ was a bit bigger than $\overline{x}_B$ and that sometimes $\overline{x}_B$ was a bit bigger than $\overline{x}_A$. Thus $\overline{x}_A - \overline{x}_B$ would sometimes be a bit bigger than 0 and sometimes be a bit smaller. In other words, $\overline{x}_A - \overline{x}_B$ has a sampling distribution. In Unit 7 we explained that, provided the samples are large enough, this sampling distribution of $\overline{x}_A - \overline{x}_B$ is approximately normal with a mean value of 0. The standard deviation of this sampling distribution is called the standard error (SE). The value of this standard error depends on the two population standard deviations and also on the two sample sizes.
3. The test statistic z is calculated as follows:

   $$z = \frac{\overline{x}_A - \overline{x}_B}{\text{SE}}.$$

   If the null hypothesis were true, then the sampling distribution of the test statistic z would be the standard normal distribution.
4. Unfortunately, the standard error is not usually known, because the population standard deviations are generally unknown – if we knew the population standard deviations, we would probably know the population means (μ_A and μ_B), and then there would be no need to test whether μ_A equals μ_B. We therefore have to estimate the standard error. In Unit 7 we calculated the following estimate of the standard error from the data being analysed:

   $$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}},$$

   where s_A and s_B are the sample standard deviations of the two samples, and n_A and n_B are the sample sizes.
5. Then the test statistic, z, is calculated as

   $$z = \frac{\overline{x}_A - \overline{x}_B}{\text{ESE}}.$$

6. The null hypothesis is rejected whenever z is far enough away from zero. For example, using a 5% significance level, the null hypothesis is rejected if
   - either $z \geq 1.96$
   - or $z \leq -1.96$.

   The figure 1.96 is the critical value.

For large samples, the procedure in stages 5 and 6 above works perfectly well, but for small samples, particularly those involving fewer than 25 values, two problems arise.

- As we mentioned in Unit 7, the sampling distribution of $\overline{x}_A - \overline{x}_B$ is approximately normal only for reasonably large sample sizes (at least 25). Thus the z-test does not necessarily work for small samples because the sampling distribution of $\overline{x}_A - \overline{x}_B$ may not then be approximately normal. This problem does not arise if the two populations in question themselves have normal distributions: in this case, the sampling distribution of $\overline{x}_A - \overline{x}_B$ is normal for all sample sizes, however large or small. Thus, if it is reasonable to

assume that the populations have normal distributions, then this difficulty is removed.

- The second problem is in the estimate of the standard error, which is

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

To be more precise, ESE is the sample estimate of the standard deviation of the sampling distribution of the difference between the sample means: $\overline{x}_A - \overline{x}_B$. It is used because the actual value of this standard deviation (i.e. the standard error) is not known. For large sample sizes, this estimate is likely to be very close to the true standard error, but, even for large samples, it is a number calculated from a sample, so it is not the same for all possible samples. Because of this variability between samples, the test statistic

$$\frac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$$

has a distribution different from the standard normal distribution that you would get if you divided $\overline{x}_A - \overline{x}_B$ by its *actual* standard deviation (i.e. the *actual* standard error) rather than this sample estimate.

If the sample sizes, n_A and n_B, are large, then ESE will vary very little from one sample to another. Its distribution will have a very small spread, and so

$$\frac{\overline{x}_A - \overline{x}_B}{\text{ESE}}$$

will have a distribution which is very close to the standard normal distribution. If either n_A or n_B is small, then this distribution will not be close to the standard normal distribution: it will be more spread out than the standard normal distribution; i.e. it will tend to have fewer values close to zero and more values away from zero. We are still assuming that the null hypothesis $\mu_A - \mu_B = 0$ is true.

This has the following consequence: if you wish to compare small samples, then, even if you think that the populations have normal distributions, you cannot use the z-test.

However, under certain circumstances another test is available: it is called Student's t-test or, simply, the ***t*-test**. This test was developed by a famous statistician, W.S. Gosset (1876–1937), who published the results of his mathematical research in this area in 1908 under the pen-name *Student*.

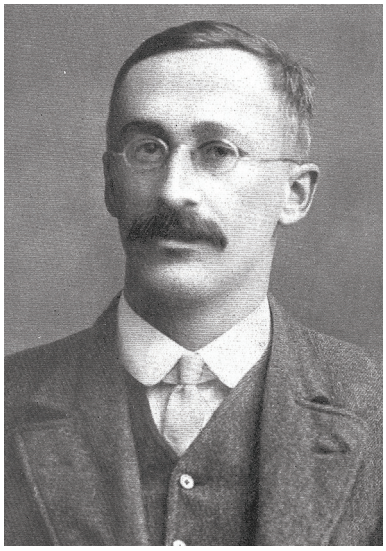
W.S. Gosset worked for the brewing company Guinness and was required by conditions of his employment to remain anonymous.

Student investigated what distribution of values you would get for an expression like

$$\frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

if you took repeated pairs of small samples from two populations whose distributions are normal with identical means and identical variances.

Since the distribution is not the standard normal distribution, the letter z can no longer be used to represent such a test statistic when it is calculated from small samples – instead, the letter t is used. The results of Student's work give a



W.S. Gosset (1876–1937)

hypothesis test that has similarities to the z-test but does not require sample sizes to be large.

3.3 The two-sample t-test

If you go on to study further statistics modules, you will probably meet many statistical techniques that have not been described in this module. Therefore, we want to show you how a hypothesis test (which you might well find in a textbook from elsewhere) can be related to the principles of hypothesis testing that you have met in the module, and also to illustrate how to do this.

If you were to look in a statistics reference book to find a hypothesis test to compare two small unrelated (i.e. unpaired) samples of numerical measurements, then you might well find something like the following summary.

Tea tasting: another t-test?

Two-sample t-test: summary

The data must be numerical measurements (such as length, weight, time) that form two unrelated samples. It is assumed that each sample is selected from a population whose distribution is normal. Also, it is assumed that the standard deviations of the two populations are equal or, equivalently, that the population variances are equal. Denoting the population means by μ_A and μ_B, the null and alternative hypotheses are:

$$H_0 : \mu_A = \mu_B \quad \text{and} \quad H_1 : \mu_A \neq \mu_B.$$

The test is carried out as follows.

1. Calculate the sample means, $\overline{x}_A$ and $\overline{x}_B$, of the two samples and the sample variances, s_A^2 and s_B^2. (s_A and s_B are the sample standard deviations.)
2. Check that the assumption of equal population variances is reasonable, or that the assumption is not seriously violated.
3. Calculate a pooled estimate s_p^2 of the **common population variance**:
$$s_\mathrm{p}^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2},$$
where n_A and n_B are the two sample sizes.
4. Calculate the test statistic:
$$t = \frac{\overline{x}_A - \overline{x}_B}{s_\mathrm{p}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}}.$$
5. The test statistic follows a ***t* distribution** with $n_A + n_B - 2$ degrees of freedom. Look up the critical value t_c of a t distribution with this number of degrees of freedom.
6. Reject H_0 in favour of H_1 if $t \geq t_\mathrm{c}$ or $t \leq -t_\mathrm{c}$. Otherwise the conclusion is that there is insufficient evidence to reject H_0.

This test is very widely used in all kinds of applications of statistics. The procedure is fairly similar to that for the two-sample z-test, but there are several important differences.

1. There is a family of t distributions. Like the family of χ^2 distributions, each member of the family has its own number of degrees of freedom. For a

two-sample t-test of unrelated samples of sizes n_A and n_B, the t distribution to use has $n_A + n_B - 2$ degrees of freedom. Hence the test statistic is compared with a critical value for that particular t distribution. (The critical values for this test, at the 5% significance level for $1, 2, \ldots, 40$ degrees of freedom, are listed in Table 2 later in this subsection.)

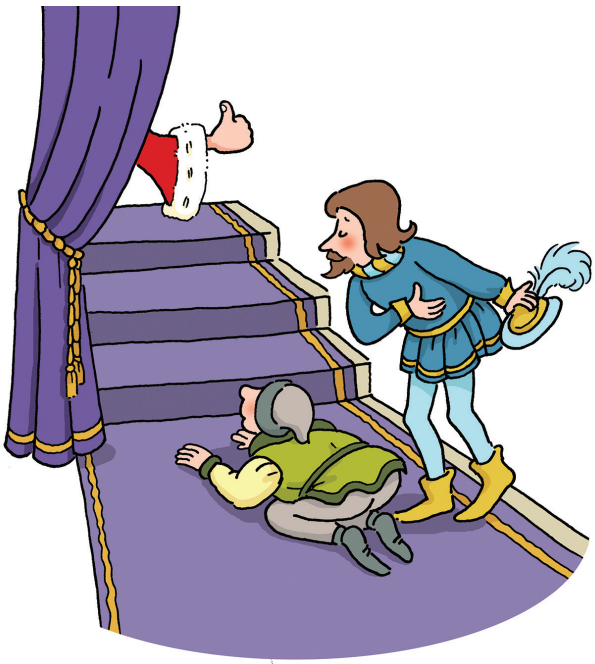
2. The t-test can be applied only if it is reasonable to assume that the populations involved have normal distributions.
3. The t-test can be applied only if it is reasonable to assume that the populations have variances that are equal.

Point 3 above gives the assumption that the population variances are equal. The validity of this assumption is readily examined as we have estimates s_A^2 and s_B^2 of the two population variances. One commonly used rule of thumb is to assume that the population variances are equal if neither one of s_A^2 and s_B^2 is greater than three times the size of the other. This is the rule we shall use in M140. If the rule is satisfied, the population variances might not be equal, but any difference between them is unlikely to be very large, and a moderate difference would seldom affect the outcome of the t-test.

Is there a common population variance?

Rule of thumb: if the sample variances (s_A^2 and s_B^2) differ by a factor of less than three, assume that there is a common population variance, or that, if the population variances differ, the difference is not large enough to invalidate the t-test.

This is not the only 'rule of thumb' of this form. Another quite common choice is to assume that the population variances are equal if neither one of s_A^2 and s_B^2 is more than *twice* the size of the other. A third choice replaces 'twice the size' with 'four times the size'. There is also a statistical hypothesis test (called the F-test) for testing the hypothesis that the population variances are equal. That test is outside the scope of M140, however you will learn in Unit 12 how to use Minitab to compare two population means without making assumptions about the population variances.



In the following example, we calculate sample variances for each of two separate samples, check that it is reasonable to assume the samples come from populations with a common population variance, and then pool the sample variances to estimate that common variance.

Example 7 Pooling sample variances

In an agricultural experiment to investigate the effect of different diets on the weights of calves, eight calves were allocated to two groups that were fed different diets, A and B. Both diets consisted of milk, hay and manufactured concentrates; the difference between them was that the concentrates in diet A were different from those in diet B. Unfortunately, one of the calves (on diet B) suffered from a disease which prevented proper digestion, so did not eat very much. That calf was therefore excluded when assessing the effects of the two diets.

The calves were kept in similar conditions, and the food intake and weight of each calf was monitored from birth. The allocation of the calves to the two groups was designed to control for birth-weight but was otherwise random. Table 1 contains some of the data collected in this experiment.

Table 1 Average daily weight gain over five weeks from birth-date (kg per day)

Calves on diet A	Calves on diet B
0.56	0.67
0.42	0.72
0.53	0.64
0.54	–

In Examples 8 and 9 we shall use the two-sample t-test to investigate whether there is a difference between the population means, μ_A and μ_B, of the daily weight gains of calves fed on the two diets. As preliminary steps towards that, we calculate the sample means and sample variances for each sample.

For Sample A,

$$\sum x_A = 0.56 + 0.42 + 0.53 + 0.54 = 2.05$$

and

$$\sum x_A^2 = 0.56^2 + 0.42^2 + 0.53^2 + 0.54^2 = 1.0625.$$

As $n_A = 4$,

$$\overline{x}_A = 2.05/4 = 0.5125$$

and

$$\sum x_A^2 - \frac{(\sum x_A)^2}{n_A} = 1.0625 - \frac{2.05^2}{4} = 0.011\,875,$$

so

$$s_A^2 = \frac{0.011\,875}{n_A - 1} = \frac{0.011\,875}{3} \simeq 0.003\,958\,3.$$

For Sample B,

$$\sum x_B = 0.67 + 0.72 + 0.64 = 2.03$$

and

$$\sum x_B^2 = 0.67^2 + 0.72^2 + 0.64^2 = 1.3769.$$

As $n_B = 3$,

$$\overline{x}_B = 2.03/3 \simeq 0.676\,666\,7$$

and

$$\sum x_B^2 - \frac{(\sum x_B)^2}{n_B} = 1.3769 - \frac{2.03^2}{3} \simeq 0.003\,266\,7,$$

so

$$s_B^2 \simeq \frac{0.003\,266\,7}{n_B - 1} = \frac{0.003\,266\,7}{2} = 0.001\,633\,35.$$

To examine whether it is reasonable to assume a common population variance, we divide the larger sample variance (in this case, s_A^2) by the smaller population variance (s_B^2):

$$\frac{s_A^2}{s_B^2} \simeq \frac{0.003\,958\,3}{0.001\,633\,35} \simeq 2.42.$$

As this ratio is less than 3, our rule of thumb says that we can pool the variances:

$$\begin{aligned} s_\mathrm{p}^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \\ &\simeq \frac{(4-1) \times 0.003\,958\,3 + (3-1) \times 0.001\,633\,35}{4 + 3 - 2} \\ &= \frac{0.015\,141\,6}{5} \simeq 0.003\,028\,3. \end{aligned}$$

Thus s_p, our estimate of the common population standard deviation, is $\sqrt{0.003\,028\,3} \simeq 0.055\,030$. It is important to record this standard deviation to at least four significant figures, as it will be used in further calculations. (You should always check that s_p^2 lies between the two sample variances, here $0.003\,958\,3$ and $0.001\,633\,35$ – you have made a calculation error if it does not.)

Example 7 is the subject of Screencast 1 for Unit 10 (see the M140 website).

Activity 10 Ball manoeuvres

Two groups of children were asked to solve a simple puzzle in which they had to manoeuvre a ball around an obstacle course and into a hole. One group, A, of children saw the obstacle course before but were not told how to negotiate it. The other group, B, of children did not see the obstacle course before but were told in advance how to negotiate it.

The length of time (in seconds) taken by each child to manoeuvre the ball round the course is shown in the table below.

Group A	Group B
2	8
7	11
8	3
3	5
5	8

(a) Calculate the sample mean and the sample variance for each group.

(b) Check whether it is reasonable to assume that the groups come from populations whose distributions have a common variance.

(c) Calculate s_p^2, the pooled estimate of the common variance of the two populations. Hence obtain a pooled estimate of the common standard deviation.

We now move to the other steps for performing a two-sample t-test. The null and alternative hypotheses are:

$$H_0 : \mu_A = \mu_B \quad \text{(the population means are equal)}$$

and

$$H_1 : \mu_A \neq \mu_B \quad \text{(the population means are not equal).}$$

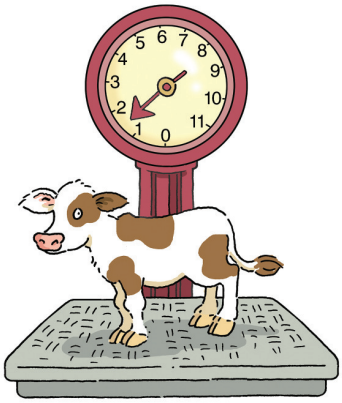
The test statistic is

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_\mathrm{p}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}}.$$

Example 8 The test statistic for weight gain of calves

For the calves experiment from Example 7,

$$t \simeq \frac{0.5125 - 0.676\,666\,7}{0.055\,030\sqrt{\dfrac{1}{4} + \dfrac{1}{3}}} \simeq \frac{-0.164\,166\,7}{0.042\,029\,9} \simeq -3.906.$$



Now, if the null hypothesis H_0: $\mu_A - \mu_B = 0$ is true, then we would expect t to be close to zero. So we ask: 'Is -3.906 close enough to zero, or should the null hypothesis be rejected in favour of the alternative hypothesis that the mean weight gains differ?' To answer this we consult a table of critical values.

Table 2 gives the 5% critical values for t distributions with different degrees of freedom. For a two-sample t-test with sample sizes n_A and n_B, the number of degrees of freedom is $n_A + n_B - 2$ (the same as the denominator in the formula for s_p^2).

Table 2 5% critical values for Student's t-test

Degrees of freedom	Critical value (t_c)	Degrees of freedom	Critical value (t_c)
1	12.706	21	2.080
2	4.303	22	2.074
3	3.182	23	2.069
4	2.776	24	2.064
5	2.571	25	2.060
6	2.447	26	2.056
7	2.365	27	2.052
8	2.306	28	2.048
9	2.262	29	2.045
10	2.228	30	2.042
11	2.201	31	2.040
12	2.179	32	2.037
13	2.160	33	2.035
14	2.145	34	2.032
15	2.131	35	2.030
16	2.120	36	2.028
17	2.110	37	2.026
18	2.101	38	2.024
19	2.093	39	2.023
20	2.086	40	2.021

(Table 2 will be referred to at various points in the unit. A copy of this table can be found in the Handbook.)

Figure 13 shows plots of t distributions for various numbers of degrees of freedom. For comparison, it also plots the normal distribution. If you look at the figure, you will see that when the number of degrees of freedom is small, the corresponding curve is comparatively low in the middle and comparatively high (i.e. further from the horizontal axis) at the extremes. As the number of degrees of freedom increases, the curves get nearer the axis at the extremes. Thus the probability of getting a large (i.e. extreme) value of t (either negative or positive) is bigger when the number of degrees of freedom is small than when it is large.

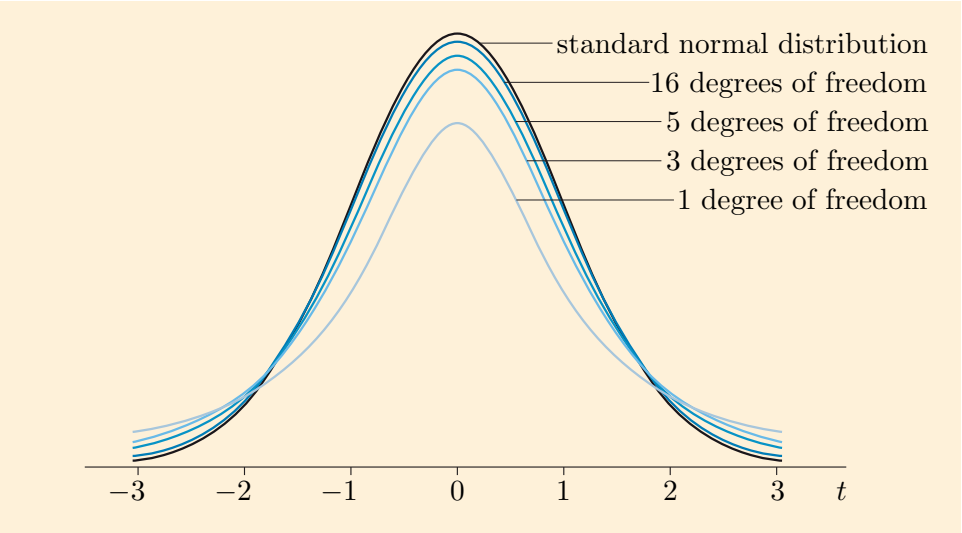


Figure 13 Sampling distribution of the test statistic t, for various numbers of degrees of freedom

Example 9 Critical value for the calves experiment

Returning to the calves experiment, here $n_A = 4$ and $n_B = 3$. Therefore the number of degrees of freedom is $4 + 3 - 2 = 5$; thus the row of Table 2 to consult is the one with 5 in the 'Degrees of freedom' column. This gives the critical value $t_c = 2.571$.

This means that if the null hypothesis is true, then the probability of obtaining a value of t greater than 2.571 is 0.025, or 2.5%, and the probability of obtaining a value of t less than -2.571 is also 0.025. So if the null hypothesis is true, then the probability of obtaining a value of t less than -2.571 or greater than 2.571 is 0.05, or 5% (see Figure 14).

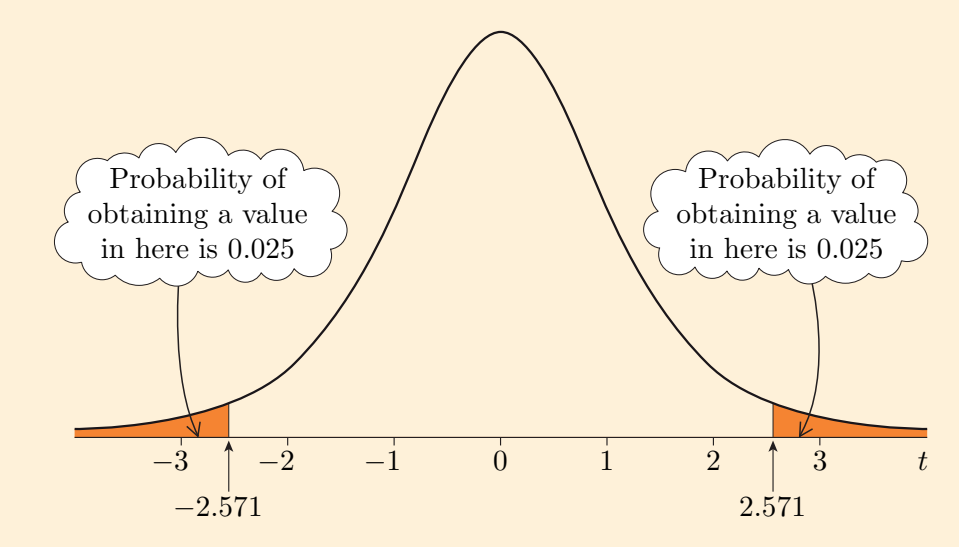


Figure 14 Sampling distribution, under the null hypothesis, of the test statistic t with five degrees of freedom

Activity 11 Conclusion from the calves experiment?

In our example, the value of t was calculated to be -3.906. Does this mean that the null hypothesis of no difference between the diets should be rejected at the 5% significance level? What do you conclude?

In general, the null hypothesis should be rejected whenever the calculated value of the test statistic t is further from zero than the critical value, i.e. whenever the value of t, ignoring any minus sign, is greater than or equal to the critical value t_c given in Table 2 for the appropriate number of degrees of freedom.

If you look at the critical values in Table 2 more closely, you may notice that as the number of degrees of freedom increases, the critical value decreases. For example, for 10 degrees of freedom the critical value is 2.228, whereas for 40 degrees of freedom it is 2.021. The reasons for this are illustrated in Figure 13: for small numbers of degrees of freedom, the tails of the distribution of the test statistic t die away less rapidly than for large numbers, so the critical value must be further from zero. This difference can be seen quite clearly (from the curves in the figure) for one or three degrees of freedom, but the curve for 16 degrees of freedom looks very similar to that of the standard normal distribution.

For a larger number of degrees of freedom, the curve (drawn at the scale of the figure) would be indistinguishable from that of the standard normal distribution. However, even these small differences in the curves produce noticeable differences in the sizes of the tails. The tails of these distributions of t all represent a larger proportion of the distribution than do the tails of the standard normal distribution. This larger proportion makes the critical value noticeably larger than 1.96: for 16 degrees of freedom it is 2.120, and for 40 degrees of freedom it is 2.021 (this is smaller than 2.120 but still larger than 1.96). Thus, as the number of degrees of freedom increases, the corresponding critical value decreases: the critical values get closer and closer to 1.96, but they never become smaller than 1.96.

Activity 12 t-test for ball manoeuvres

For the ball manoeuvres experiment in Activity 10 carry out a t-test at the 5% significance level to test the null hypothesis

$H_0 : \mu_A = \mu_B$ (seeing course or getting instruction are equivalent)

against the alternative hypothesis

$H_1 : \mu_A \neq \mu_B$ (seeing course or getting instruction are not equivalent).

Key values for a two-sample t-test

The information you need to know for a two-sample t-test is:

- the sample means ($\overline{x}_A$ and $\overline{x}_B$)
- the sample sizes (n_A and n_B)
- the sample standard deviations (s_A and s_B) or the pooled standard deviation (s_p) or the corresponding variances.

Activity 13 Plant heights

This concerns an experiment to investigate the heights of two different varieties of lupin: *Lupinus arboreus* and *Lupinus hartwegii*. Here are the summaries of the data from two samples: one of each of these varieties. All the plants were grown at the same time in similar conditions in a nursery, and the height of each (in metres) was measured on the same day.

Lupinus arboreus:

- sample size $n_A = 5$
- mean $\overline{x}_A = 1.252$
- standard deviation $s_A = 0.051$.

Lupinus hartwegii:

- sample size $n_B = 6$
- mean $\overline{x}_B = 1.023$
- standard deviation $s_B = 0.038$.

Carry out a t-test on these samples to investigate whether there is a significant difference between the heights of the two varieties. (With summary data like these, you cannot check whether the heights are normally distributed – but assume that they can be.)

(a)

(b)

Examples of (a) *Lupinus arboreus* and (b) *Lupinus hartwegii*

Before asking you to analyse the data from the experiment that you are conducting, we will say a little more about the connection between two-sample z-tests and two-sample t-tests, to help show the rationale behind the latter.

The test statistic of the two-sample z-test is

$$z = \frac{\overline{x}_A - \overline{x}_B}{\mathsf{SE}},$$

where

$$\mathsf{SE} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}.$$

When the t-test is applicable, the two population variances are equal and s_p^2 is an estimate of that common variance. Substituting s_p^2 for both σ_A^2 and σ_B^2 (which are generally unknown) gives

$$\mathsf{SE} = \sqrt{\frac{s_\mathsf{p}^2}{n_A} + \frac{s_\mathsf{p}^2}{n_B}} = s_\mathsf{p}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

Then the above test statistic for the two sample z-test becomes

$$\frac{\overline{x}_A - \overline{x}_B}{s_\mathsf{p}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}},$$

which is the test statistic for the two-sample t-test.

You have now covered the material related to Screencast 2 for Unit 10 (see the M140 website).

3.4 Analysis of mustard seedling data

3. Analyse

The analysis of your mustard seedling data will form part of the tutor-marked assignment covering this unit. When your mustard seedlings have grown sufficiently, you should collect data on them, following the instructions in Subsection 2.6. Using these data, you will be able to test whether light affects root growth.

Even if you have not got 10 measurements from each group, you can still carry out the test – provided you have at least two measurements from each group!

You should have 10 measurements from seedlings grown in the light (Group A) and 10 from seedlings grown in the dark (Group B): all from seedlings whose

stems you cut. Before carrying out the t-test on these data, you should consider whether you can assume that the populations satisfy the necessary distributional conditions. The only way you can do this is by examining the data that you have collected. We shall now demonstrate how to do this using the following data, which we hope are not too different from yours.

In Table 3, measurements which are bold were obtained from seedlings whose stems were cut during their growth. A cross indicates a seed which did not germinate.

Table 3 Lengths of roots (in mm) obtained in a mustard seedlings experiment

Seedlings grown in light					Seedlings grown in dark				
21	**39**	27	**31**	×	**22**	×	21	×	39
×	21	26	**13**	12	20	×	**16**	**20**	×
52	**39**	×	11	**55**	**14**	**32**	**28**	×	**36**
50	×	8	**29**	**17**	24	**41**	20	**17**	**22**

Activity 14 Stemplots of seedling data

Prepare separate stemplots of the lengths of the roots for each of the two samples of 10 cut seedlings from the given data (i.e. of the values which are bold).

Activity 15 Sample variances of seedling data

Calculate the sample variance of root-length for the sample of seedlings grown in the light. Do the same for the sample grown in the dark. Can we treat the samples as coming from populations with equal variances?

From the stemplots, the two samples of data each look as if they could have come from populations with a normal distribution. Also, our rule of thumb says we may treat the population variances (and hence also their standard deviations) as being equal. (Also, the spreads of the observations seem similar in the two stemplots, suggesting the populations have similar variances.) With such small samples it is not possible to be more precise than this, but there is certainly no strong evidence that the distributional conditions required for the t-test are not satisfied. Nor, again because they are so small, would there be any such evidence even if the samples were less symmetric. Thus Activities 14 and 15, together with a large amount of similar data about living things collected by scientists, suggest that we can apply the t-test to these samples of data.

The following summarises the procedure for calculating the test statistic.

Calculation of the test statistic for the two-sample t-test

Denote the two samples by A and B, and their sizes by n_A and n_B.

1. Calculate $\sum x_A$, $\sum x_B$, $\sum x_A^2$ and $\sum x_B^2$.
2. Calculate the sample means $\overline{x}_A = \sum x_A / n_A$ and $\overline{x}_B = \sum x_A / n_A$.
3. Calculate

$$\sum x_A^2 - \frac{(\sum x_A)^2}{n_A}$$

and divide it by $n_A - 1$ to obtain s_A^2. Similarly, divide

$$\sum x_B^2 - \frac{(\sum x_B)^2}{n_B}$$

by $n_B - 1$ to obtain s_B^2.

4. Divide the larger of s_A^2 and s_B^2 by the smaller to check whether it is less than 3. If this is the case, it is sensible to assume a common population variance.
5. Calculate a pooled estimate s_p^2 of the common population variance:
$$s_\mathrm{p}^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}.$$
6. Calculate the test statistic:
$$t = \frac{\overline{x}_A - \overline{x}_B}{s_\mathrm{p}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}}.$$
Round this to three decimal places.

After completing this section and collecting your seedling data, you are in a position to analyse the results of your experiment!

Exercises on Section 3

Exercise 3 Hay and barley

Another experiment on calves, similar to that in Example 7, was carried out to compare a diet H including hay with a diet B including barley straw. The average weight gain (kg per day) over five weeks after the birth-date was again measured for each calf. The results for the two diets are summarised here.

Hay:

- sample size $n_H = 20$
- mean $\overline{x}_H = 0.542$
- standard deviation $s_H = 0.081$.

Barley straw:

- sample size $n_B = 19$
- mean $\overline{x}_B = 0.554$
- standard deviation $s_B = 0.088$.

Carry out a t-test on these sample data to test whether the two diets differ in their effect on average weight gain. (Assume that the weight gains of calves are normally distributed.)

Exercise 4 Comparing production lines

A manufacturer wishes to compare the performance of two biscuit production lines, A and B. The lines produce packets of biscuits with a nominal weight of 300 grams. Two random samples of 15 packets from each of the two lines are weighed (in grams). The sample data are summarised as follows.

Line A:

- sample size $n_A = 15$
- mean $\overline{x}_A = 309.8$
- standard deviation $s_A = 3.58$.

Line B:

- sample size $n_B = 15$
- mean $\overline{x}_B = 305.2$
- standard deviation $s_B = 4.73$.

Carry out a t-test on these sample data to test whether the two production lines produce packets of different average weight. (Assume that the weights of packets of biscuits produced by each production line are normally distributed.)

4 The t-test for one sample and matched-pairs samples

In this section we shall introduce a version of the t-test that can be used to analyse a single sample of suitable data when the sample size is not large. We then show how this test can be adapted to produce a useful test for data from experiments involving matched pairs. The method is almost identical to the one-sample z-test that we use when the sample size is large and the population standard deviation is unknown. (That test was the topic of Subsection 5.2 of Unit 7.) The test statistic is calculated in the same way, but now we compare it with the tabulated critical values of a t distribution, rather than the critical values of a normal distribution.

4.1 The one-sample t-test

The t-test described in Subsections 3.2 and 3.3 is used to compare two unrelated (i.e. unpaired) samples. We pointed out there that in many ways the procedure is similar to the z-test for the difference between two population means. You may have wondered whether there is a one-sample t-test that resembles the one-sample z-test, but which can be used with a small sample of data. The following example will be used to address this question.

Example 10 A small sample of tomato plants

A tomato grower decides to try out a new fertiliser on one variety of outdoor bush tomato plants that he grows. Previously this variety has produced an average yield of 4 kg of tomatoes per plant. The grower wants to investigate whether this average yield would change if he switched to the new fertiliser. He has room to experiment with only five plants on the new fertiliser. The yields of tomatoes from each of these five plants, in kg, are:

$$3.6 \quad 3.2 \quad 3.1 \quad 2.6 \quad 3.9$$

If the population mean of the yield (in kg per plant) using this new fertiliser is denoted by μ, then the grower's null hypothesis is that μ is the same as the average yield used to be. In symbols:

$$H_0 : \mu = 4.$$

His alternative hypothesis is that the new fertiliser changes the average yield. In symbols:

$$H_1 : \mu \neq 4.$$

If the sample size had been large in Example 10, then the grower could have used the one-sample z-test; however, the sample is certainly not large enough. He could use the sign test, provided he changed his hypothesis to refer to the population median rather than the mean. However, these data are measurements, and the sign test takes account only of whether each data value is above or below a particular number. Thus the sign test does not use all the information in the data. It is possible to use much more of this information by using a t-test. It must be stressed, though, that this can be done only if it is reasonable to assume that the population distribution is normal. Our tomato grower might well feel that this assumption is justified. (Many measurements of living things are normally distributed. For example, the heights of adult men closely follow a normal distribution.)

Activity 16 Key values from the tomato experiment

The key values are the sample size n, sample mean $\overline{x}$ and sample standard deviation s. State the sample size and calculate the sample mean and sample standard deviation for the tomato grower's data.

Activity 17 Could the z-test be used?

(a) We want to test the hypothesis H_0: $\mu = A$, where $A = 4$, for the tomato grower's experiment. What test statistic would you use if you were trying to apply a z-test to his data?

(b) Why is the z-test not appropriate here?

In a one-sample t-test, the null hypothesis is $H_0 : \mu = A$ and the alternative hypothesis is $H_1 : \mu \neq A$. These are precisely the hypotheses used in a one-sample z-test. The test statistic for this z-test is

$$z = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}},$$

as noted in the solution to Activity 17. Replacing z by t gives the test statistic for the one-sample t-test.

The test statistic for the one-sample t-test:

$$t = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

Example 11 t-test statistic for the tomato experiment

For the tomato grower's experiment, we found in Activity 16 that $n = 5$, $\overline{x} = 3.28$ and $s = 0.247$. Also $A = 4$, as the null hypothesis is H_0: $\mu = 4$. Hence ESE $\simeq 0.496\,99/\sqrt{5} \simeq 0.222\,26$, and the test statistic for a t-test is:

$$t = \frac{\overline{x} - 4}{\text{ESE}} = \frac{3.28 - 4}{0.222\,26} \simeq -3.239.$$

This, of course, is precisely the value found in part (a) of Activity 17.

If H_0 were true, then the population mean would be 4, so the sample mean $\overline{x}$ would probably be close to 4. Hence $\overline{x} - 4$ would be near 0, and so t would be near 0. Thus, as with the z-test, values of t sufficiently far from 0 (positive or negative) result in rejection of the null hypothesis. To know whether to reject the null hypothesis, we need a critical value – you will probably not be surprised to learn that this comes from the table of critical values in Table 2 (Subsection 3.3). To use this table we need to know what number of degrees of freedom to look up. For the one-sample t-test the rule is:

For a sample of size n, the number of degrees of freedom is $n - 1$.

Activity 18 Critical value for the tomato experiment

Find the critical value at the 5% significance level for the tomato grower's test.

Example 12 Comparing the test statistic to the critical value

The rejection rule for the tomato grower is therefore: reject H_0 in favour of H_1 if the sample value of the test statistic t is greater than or equal to 2.776 or less than or equal to -2.776. Since $t = -3.239$, which is less than -2.776, the tomato grower should reject H_0 and conclude that, on the basis of his sample, the average yield with the new fertiliser is not 4 kg per plant. Of course he may have some reservations about this conclusion. Perhaps this year the weather was bad for tomato plants, for example.

Example 13 Using a sign test for the tomato experiment

In order to compare the t-test with the sign test, it is informative to re-examine the data on tomato plants using the sign test. The null hypothesis would be

H_0 : The median yield per plant is 4,

and the alternative hypothesis would be

H_1 : The median yield per plant is not 4.

The yields of the five plants were 3.6, 3.2, 3.1, 2.6 and 3.9, which are all less than 4, so the value of the test statistic is 0.

From Table 8 in Subsection 4.1 of Unit 6 (and repeated in the Handbook), we find that there is *no* value of the test statistic for which we would reject H_0 at the 5% significance level. Hence we would not reject it here.

In Example 12, the t-test rejected the null hypothesis. However, in Example 13, the sign test did not reject the null hypothesis. This is because the sign test uses less information than the t-test. In general, the t-test is more likely than the sign test to reject a null hypothesis that is false. (That is, the t-test is less likely to make a Type 2 error.) We say that the t-test is more **powerful** than the sign test.

Power of the t-test and the sign test

The t-test is said to be a more powerful test than the sign test because the t-test is better at identifying a null hypothesis that is false.

The price that is paid in using the t-test (rather than the sign test) is that an extra assumption is needed – we must assume that the sample is from a normal distribution. If this assumption is reasonably close to reality, then it is much better to use the t-test.

Here is a summary of the procedure for the one-sample t-test.

Procedure: the one-sample t-test

The test applies to a sample of data consisting of numerical measurements, when the population from which the data comes can be assumed to have a normal distribution.

1. Denoting the population mean by μ, the null and alternative hypotheses are:
$$H_0 : \mu = A$$
$$H_1 : \mu \neq A.$$
2. Calculate the sample mean, $\overline{x}$, and the sample standard deviation, s.
3. Calculate the estimated standard error:
$$\text{ESE} = \frac{s}{\sqrt{n}},$$
where n is the sample size.
4. The test statistic is
$$t = \frac{\overline{x} - A}{\text{ESE}}.$$
5. The critical value at the 5% significance level is the value of t_c for $n - 1$ degrees of freedom in Table 2 (Subsection 3.3).
6. Reject H_0 in favour of H_1 at the 5% significance level if
   - either $t \geq t_c$
   - or $t \leq -t_c$.

   Otherwise H_0 is not rejected at the 5% significance level.
7. State the conclusion that can be drawn from the test.

You have now covered the material related to Screencast 3 for Unit 10 (see the M140 website).

4.2 The matched-pairs t-test

In various experiments, **matched pairs** are used to remove the effects of factors that would otherwise be uncontrollable. For example, suppose a shoe manufacture wants to test which of two materials makes heels that last longer. One approach would be to make some pairs of shoes with one material and some with the other. Then let people wear the shoes for two months, after which the wear on the heels of the shoes would be measured. This would answer the question if enough pairs of shoes were used in the experiment. However, differences in wear would reflect not just differences in the materials, but also differences in the amount shoes were used, differences in weights of the people, and so forth.

One way of reducing these latter effects would be to make one heel of a pair of shoes from one material and make the other heel *of that same pair of shoes* from the other material. Then the difference in wear between the left and right shoes of a pair would not reflect differences in usage (assuming the wearer did not hop a lot), nor differences in a wearer's weight, as the same person wore both. In this form of experiment, each pair of shoes would give a matched pair of measurements, and the differences in wear between the left and right heels of a pair of shoes would be the data that are analysed.

Sometimes matching isn't with the same person. Each person in a group is *matched* as closely as possible with a person in another group in terms of age and any other aspects that we might want to control for.

Pairing

In a matched-pairs experiment, items are paired in such a way that the factor of interest (but little else) differs between the two items that form a pair. The statistical analysis is then based on the differences between items within a pair.

Taking differences combines two measurements into one, so that methods for analysing a single sample become appropriate. If the number of pairs is large (over 25), then a one-sample z-test might be used. For smaller samples, a t-test should be used if it can be assumed that the population of differences has a normal distribution.

To illustrate this use of the one-sample t-test, we shall look at some data that come from a paper by Cushny and Peebles, published in 1904. (These data were analysed by Gosset (Student) in the 1908 paper in which he published the results that led to the t-test.) These researchers wanted to investigate whether two different forms, L and R, of a drug, hyoscyamine hydrobromide, differ in their capacity to induce sleep. (Note that hyoscyamine hydrobromide is not now commonly used as a sleep-inducing drug.) They conducted an experiment with ten patients. Five of the patients received form R of the drug first, and their gain in sleep was recorded. After a suitable time had elapsed they were given form L instead and the same measurements recorded. The other five patients received the two drugs in the opposite order.

Activity 19 Benefits of the experimental design

The experiment was designed to reduce the effects of two sources of variation that could affect results. What effects were reduced?

The results from the experiment are given in Table 4. A positive figure means that the patient got more sleep with the drug than without; a negative figure means that he or she got less sleep with the drug.

Table 4 Sleep gained (hours) by the use of hyoscyamine hydrobromide

Patient	Form L	Form R
1	+1.9	+0.7
2	+0.8	−1.6
3	+1.1	−0.2
4	+0.1	−1.2
5	−0.1	−0.1
6	+4.4	+3.4
7	+5.5	+3.7
8	+1.6	+0.8
9	+4.6	0.0
10	+3.4	+2.0

Note that patients vary considerably in their responsiveness to both drugs. For example, patient 7 is very responsive to both drugs, whilst patient 5 is much less affected.

It would not be appropriate to analyse these data using the two-sample t-test from Section 3 because the data are in matched pairs: there is a pair of sleep-gain measurements for each patient. Making effective use of the matching leads to a more powerful test.

To use the matched-pairs t-test on these data, first we find the difference, for each patient, between the hours of sleep gained using form L and the hours gained using form R.

Activity 20 Forming differences

For each patient in Table 4, calculate the difference, d, in hours of sleep gained: $L - R$.

The null hypothesis is that the two forms of the drug are equally effective: i.e. on average, it does not make any difference which form a patient receives. If we denote the population mean of the hours of sleep gained with form L by μ_L, and the population mean for form R by μ_R, then the null hypothesis is

$$H_0 : \mu_L = \mu_R \quad \text{or} \quad H_0 : \mu_L - \mu_R = 0.$$

If the population mean of the difference between the hours of sleep gained with form L and with form R is denoted by μ_d, then $\mu_d = \mu_L - \mu_R$.

Now the null hypothesis is just that the population mean difference is zero:

$$H_0 : \mu_d = 0,$$

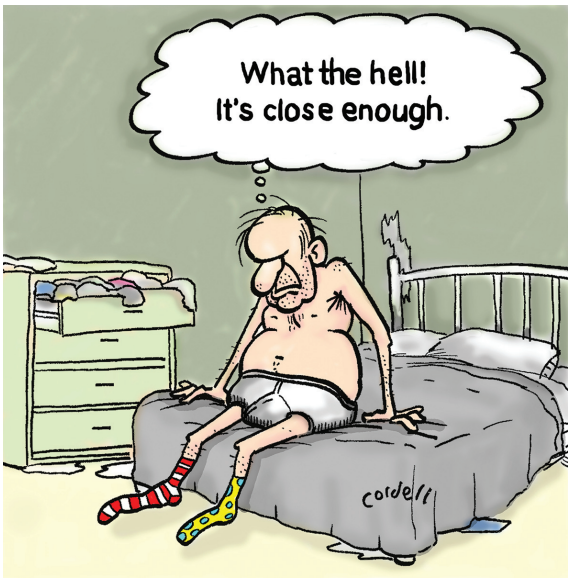
and the alternative hypothesis is

$$H_1 : \mu_d \neq 0.$$

We now have a single sample of 10 numbers, the *differences*, and null and alternative hypotheses about the population mean *difference* of the population from which this sample of 10 *differences* comes. If it is reasonable to assume that the distribution of this population of differences is normal, then we can carry out a one-sample t-test on these differences.

Activity 21 Testing for a mean difference of 0

Carry out a one-sample t-test on the differences, using d instead of x in the formulas. What do you conclude?



An unmatched pair

You have now covered the material related to Screencast 4 for Unit 10 (see the M140 website).

The following summarises the procedure for a matched-pairs t-test.

Procedure: the matched-pairs t-test

1. Calculate the differences between the two values in each pair.
2. The null and alternative hypotheses are
   $$H_0 : \mu_A = \mu_B$$
   $$H_1 : \mu_A \neq \mu_B,$$
   where μ_A and μ_B are the population means of the two populations involved.

   Replace these by the equivalent hypotheses
   $$H_0 : \mu_d = 0$$
   $$H_1 : \mu_d \neq 0,$$
   where μ_d is the population mean of the population of differences between the matched pairs.
3. If it can be assumed that this population of differences has a normal distribution, then apply the one-sample t-test with $A = 0$ and d instead of x in the formulas, to the sample of differences.

Activity 22 Comparing weighing machines

A scientist has two pieces of weighing equipment, A and B, in her laboratory, both of which she suspects may be inaccurate, though she does not know in

what way. She decides to begin an investigation into their accuracy by comparing their readings for the weights of different objects. She weighs nine objects on both pieces of equipment, with the following results:

	Weight in grams	
Object	Equipment A	Equipment B
1	3.6	3.3
2	4.3	4.4
3	11.4	11.2
4	15.9	15.5
5	16.4	16.6
6	18.7	18.7
7	21.1	20.7
8	21.8	21.4
9	24.1	23.8

The scientist wants to test whether one piece of equipment gives higher weights than the other.

(a) For each object, calculate the weight given by A minus the weight given by B. Calculate the mean and sample standard deviation of these differences.

(b) Give the hypothesis to be tested and the alternative hypothesis.

(c) Calculate the test statistic.

(d) State the number of degrees of freedom and give the critical value for the test.

(e) What is the result of the hypothesis test?

(f) State your conclusion.

Exercises on Section 4

Exercise 5 One-sample t-test

A sample of 12 items has a sample mean of 8.2 and a sample variance of 3.7. Using a one sample t-test (where in each case μ is the mean of the population from which the sample was taken):

(a) test the null hypothesis H_0: $\mu = 10$ against the alternative hypothesis H_1: $\mu \neq 10$;

(b) test the null hypothesis H_0: $\mu = 7.5$ against the alternative hypothesis H_1: $\mu \neq 7.5$.

Exercise 6 Drug comparison

Use the matched-pairs t-test to analyse data on the effect of a new drug on the weight of male patients. The data on the five matched pairs of males are given below.

Experiment group (new drug taken)		Control group (old drug taken)		
Patient number	Weight change (N)	Patient number	Weight change (O)	Difference ($N - O$)
8	-7	1	-2	-5
16	-7	2	+1	-8
20	-3	4	-3	0
13	-1	10	+1	-2
17	-2	14	-5	+3

Is there a significant difference between the effect of the new drug and that of the old drug?

5 Confidence intervals from t-tests

As with other hypothesis tests, it is often important to go beyond the test and obtain estimates in the form of confidence intervals. For example, the tomato grower in Subsection 4.1 would probably like an estimate of the yield of tomatoes which he is likely to obtain from the new fertiliser. The new fertiliser might be cheaper or easier to apply than his old fertiliser, and an interval estimate could help him to decide whether it was worth using the new fertiliser, even though it did not match up to the old fertiliser.

There is an underlying structure to many forms of confidence interval. It is common to both confidence intervals for a single population mean and confidence intervals for comparing two population means.

Confidence interval for a mean or the difference between two means

The lower limit of the confidence interval is

$$\text{point estimate} - (z \text{ or } t \text{ critical value}) \times \text{ESE},$$

and the upper limit is

$$\text{point estimate} + (z \text{ or } t \text{ critical value}) \times \text{ESE},$$

where ESE is the estimated standard error of the point estimate.

(As noted in Subsection 4.1 of Unit 9, an estimate that consists of a single number (rather than a range of values) is called a point estimate.)

A critical value of z is used if the standard error is known or estimated from a large sample (or samples, when there are two populations). The critical value of a t distribution is used if the standard error is unknown and is estimated for a small sample (or samples).

For z, we use the 5% critical value to construct a 95% confidence interval, and the 1% critical value to construct a 99% confidence interval. For t distributions, only 95% confidence intervals will be constructed by hand in M140, as Table 2 (Subsection 3.3) only gives 5% critical values.

5.1 Confidence intervals from one sample and matched-pairs t-tests

Let us suppose the tomato grower wants a confidence interval for μ, the mean yield (in kg per plant) of tomatoes that he will obtain with the new fertiliser. If he were to calculate a confidence interval based on the z-test, then he would calculate the sample mean $\overline{x}$ and the sample standard deviation s. Then the sample estimate of the standard error is

$$\text{ESE} = \frac{s}{\sqrt{n}},$$

so the 95% confidence interval for μ would be

$$(\overline{x} - 1.96 \times \text{ESE}, \overline{x} + 1.96 \times \text{ESE}).$$

Recall that 1.96 is the critical value for the z-test. However, in the hypothesis test he had to use a t-test rather than a z-test because his sample is small. For this same reason he must also calculate his confidence interval differently – it must be based on the t-test rather than the z-test.

The calculation is almost exactly the same as for the interval based on the z-test, but for one modification: he must use the critical value t_c from Table 2 (Subsection 3.3). This is the same critical value as the one which he used in the hypothesis test in Subsection 4.1: that for $n - 1$, i.e. 4, degrees of freedom.

Thus the 95% confidence interval for μ is

$$(\overline{x} - t_c \times \text{ESE}, \overline{x} + t_c \times \text{ESE}),$$

where ESE is still $s/\sqrt{n}$.

From Activities 16 and 18, $\overline{x} = 3.28$, $s \simeq 0.496\,99$, $n = 5$ and $t_c = 2.776$. So ESE $\simeq 0.496\,99/\sqrt{5} \simeq 0.222\,26$ (as in Example 11). Thus

$$t_c \times \text{ESE} \simeq 2.776 \times 0.222\,26 \simeq 0.62,$$

rounded to the same level of accuracy as the sample mean. So the tomato grower's 95% confidence interval for the mean yield, in kg per plant, is $(3.28 - 0.62, 3.28 + 0.62)$, which equals $(2.66, 3.90)$.

We can summarise this method as follows.

95% confidence interval for the population mean of a normally distributed population

The confidence interval is $\left(\overline{x} - t_c\frac{s}{\sqrt{n}}, \overline{x} + t_c\frac{s}{\sqrt{n}}\right)$,
where n, $\overline{x}$, t_c and s are as in the procedure for the one-sample t-test. Thus t_c is the critical value from Table 2 for $n - 1$ degrees of freedom.

The same argument also applies to the confidence interval from the matched-pairs t-test. So for such intervals, the formula is the same as above but with $\overline{d}$ instead of $\overline{x}$; that is, $\left(\overline{d} - t_c\, s/\sqrt{n},\ \overline{d} + t_c\, s/\sqrt{n}\right)$.

Activity 23 An enthusiastic gardener

An enthusiastic gardener wished to investigate whether sunflowers planted in her garden would indeed grow to a height of 4 metres, as claimed on the seed packet. She collected data on the heights, in metres, of a sample of

15 sunflowers grown in her garden. The results are summarised here.

$$n = 15, \quad \text{sample mean } \overline{x} = 3.60, \quad \text{sample standard deviation } s = 1.18.$$

Calculate a 95% confidence interval for the population mean of sunflowers grown in her garden.

Activity 24 Confidence interval from a paired t-test

In Subsection 4.2 we examined data on the sleep gain of two forms of a drug, form L and form R. We set $\mu_d = \mu_L - \mu_R$ and in Activity 21 we tested the null hypothesis that $\mu_d = 0$.

Use the solution to that activity to find a 95% confidence interval for μ_d.

As $\mu_d = \mu_L - \mu_R$, the last activity determined a confidence interval for $\mu_L - \mu_R$. In words, μ_d is the population mean of the difference between the hours of sleep gained by the same patient using form L and using form R of the drug hyoscyamine hydrobromide. Thus you have calculated a 95% confidence interval for the mean difference between the hours of sleep gained using the two forms of the drug.

5.2 Confidence intervals from two unrelated samples

The confidence interval for the difference between two population means can be based on critical values of z when we have a large sample from each population. When population standard deviations are unknown, this should not be done if one or both samples are small. Instead, a confidence interval can be based on the t distribution, provided the same conditions are satisfied that are required in the t-test for two unrelated samples. These conditions are:

- Two unrelated samples of independent observations are taken, one sample from each of the two populations of interest.
- The distributions of the two populations are normal, and their standard deviations are equal.

The confidence interval is given by the general form

$$\text{lower limit} = \text{point estimate} - t_c \times \text{ESE}$$

and

$$\text{upper limit} = \text{point estimate} + t_c \times \text{ESE}.$$

To apply this formula, denote the sample sizes by n_A and n_B, the sample means by $\overline{x}_A$ and $\overline{x}_B$, and the sample variances by s_A^2 and s_B^2. The pooled estimate of the common variance is calculated as before:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2},$$

and then

$$\text{ESE} = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

As with the corresponding hypothesis test, the appropriate t distribution has $n_A + n_B - 2$ degrees of freedom, and t_c is obtained from Table 2 (Subsection 3.3). Thus the formula for the confidence interval is as follows.

95% confidence interval for the difference between the population means of two unrelated normally distributed populations with equal standard deviations

The confidence interval is

$$\left((\overline{x}_A - \overline{x}_B) - t_c s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}},\ (\overline{x}_A - \overline{x}_B) + t_c s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}\right),$$

where t_c is the critical value from Table 2 for $n_A + n_B - 2$ degrees of freedom.

Example 14 Confidence interval for calves' weight gains

The first example in Subsection 3.3 concerned the weight gains of calves fed two different diets, diet A and diet B. We will calculate a 95% confidence interval for the difference in average weight gain (kg per day) on the two diets. Useful summary statistics for these data were obtained in Examples 7 and 9:

$$n_A = 4, \quad n_B = 3, \quad \overline{x}_A = 0.5125, \quad \overline{x}_B \simeq 0.676\,666\,7,$$
$$s_p \simeq 0.055\,030, \quad t_c = 2.571.$$

To apply the above formula, we first calculate

$$t_c s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \simeq 2.571 \times 0.055\,030 \sqrt{\frac{1}{4} + \frac{1}{3}} \simeq 0.108$$

and

$$\overline{x}_A - \overline{x}_B = 0.5125 - 0.676\,666\,7 \simeq -0.164.$$

Thus the confidence interval for the mean weight gain, in kg per day, is $(-0.164 - 0.108, -0.164 + 0.108)$, which is $(-0.272, -0.056)$.

When calculating a confidence interval for two unrelated samples, we make the assumption that the two populations have variances that are equal. Checking that this assumption was reasonable was unnecessary in Example 14, as we had checked it for our two populations of calves before testing the hypothesis $\mu_A = \mu_B$ in Subsection 3.3. This will be the case whenever construction of the confidence interval is preceded by a hypothesis test. When a hypothesis test has not been carried out, though, the assumption must be examined, using the same rule of thumb as before.

Checking equality of variances

The assumption of equal population variances holds acceptably well if the ratio of the larger sample variance to the smaller sample variance is less than 3. This condition should be checked before forming a confidence interval between two population means on the basis of two unrelated samples, unless the condition has already been checked in the course of a hypothesis test.

Activity 25 Sun and shade

The data on heights of sunflowers collected in Activity 23 were for sunflowers grown in a sunny flower bed in the garden, bed A. Sunflowers were also grown in a shady flower bed of the same garden, bed B. The gardener wanted an interval estimate of difference in heights for sunflowers grown in a sunny flower bed compared with sunflowers grown in a shady flower bed.

Results are summarised here.

Sunny:

- sample size $n_A = 15$
- mean $\overline{x}_A = 3.60$
- standard deviation $s_A = 1.18$.

Shady:

- sample size $n_B = 20$
- mean $\overline{x}_B = 2.76$
- standard deviation $s_B = 1.09$.

Check that the assumption of equal population variances holds acceptably well, and calculate a 95% confidence interval for the mean difference between the heights of sunflowers grown in the shady flower bed and those grown in the sunny flower bed.

You have now covered the material related to Screencast 5 for Unit 10 (see the M140 website).

Exercises on Section 5

Exercise 7 Confidence interval for equipment

In Activity 22 (Subsection 4.2) you tested a hypothesis about weight readings given by Equipment A and Equipment B. Let μ_d denote the average difference in weight that they give. Using results obtained in Activity 22, form a 95% confidence interval for μ_d.

Exercise 8 Confidence interval for production lines

Exercise 4 (Section 4) concerned the weights of packets of biscuits produced on two production lines. Using results obtained in that exercise, calculate a 95% confidence interval for the difference $\mu_A - \mu_B$, where μ_A and μ_B are the mean weights of packets of biscuits from the two production lines.

6 One-sided alternative hypotheses

Until now, when using z-tests and t-tests we have rejected the null hypothesis if the test statistic is larger in size (positive or negative) than a critical value. Such tests are sometimes referred to as **two-sided hypothesis tests**, and they are much the most common form of hypothesis test.

In some situations, however, we only want to reject the null hypothesis if the test statistic is large and positive, while in other situations, we only want to reject the null hypothesis if the test statistic is large and negative. In both these cases, the alternative hypothesis specifies a direction and is said to be a **one-sided alternative hypothesis**. For a one-sample z-test or t-test, where the null hypothesis is H_0: $\mu = A$, a one-sided alternative hypothesis will have the form

$$H_1 : \mu < A \quad \text{or} \quad H_1 : \mu > A.$$

Similarly, for a two-sample z-test or t-test, the null hypothesis is H_0: $\mu_A = \mu_B$, while the one-sided alternative hypothesis is

$$H_1 : \mu_A < \mu_B \quad \text{or} \quad H_1 : \mu_A > \mu_B.$$

A hypothesis test that uses a one-sided alternative hypothesis is said to be a **one-sided test**.

As an example of where a one-sided alternative hypothesis would be appropriate, suppose a student takes a multiple choice test in which each question offers a choice of four answers, of which only one is correct. The null hypothesis might be 'the student just guesses', while the alternative is 'the student knows something'. If the null hypothesis is true, then the student should score about 25%. If he scores much more, we would decide that he was not just guessing, so that the alternative hypothesis is true. If he scores much less than 25%, then we would not favour the alternative hypothesis. Rather, we would conclude that the student was just guessing (so the null hypothesis is true) *and* he was also unlucky! If μ denotes the proportion of questions that the student will get right on average if he takes versions of the test many times, then appropriate hypotheses would be

$$H_0 : \mu = 0.25 \quad \text{and} \quad H_1 : \mu > 0.25,$$

as μ should be more than 0.25 if the student knows something.

One-sided z-tests and t-tests are performed in exactly the same way as the corresponding two-sided tests, except:

- the critical value changes
- we only consider rejecting the null hypothesis if the value of the test statistic is in the direction specified by the alternative hypothesis.

Here we will only give further detail for t-tests, as this adequately demonstrates the similarities and minor differences between one-sided and two-sided hypothesis tests. Table 5, after the following boxes, gives critical values for one-sided t-tests at the 5% significance level. The number of degrees of freedom are the same as before: $n - 1$ for a one-sample test and $n_A + n_B - 2$ for a two-sample test.

"We prefer to call this test 'multiple choice', not 'multiple guess'."

One-sided t-test for one sample or matched-pairs samples

The null hypothesis is again H_0: $\mu = A$. The test statistic t is calculated as for the two-sided one-sample/matched-pairs t-test:

$$t = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

The critical value at the 5% significance level is the value of t_c for $n - 1$ degrees of freedom in Table 5.

If the alternative hypothesis is H_1: $\mu > A$, we reject H_0 at the 5% significance level if $t > t_c$.

If the alternative hypothesis is H_1: $\mu < A$, we reject H_0 at the 5% significance level if $t < -t_c$.

One-sided two-sample t-test

The null hypothesis is again H_0: $\mu_A = \mu_B$. As with the two-sided two-sample test, check that it is reasonable to assume equal variances, and, assuming it is, calculate the test statistic:

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}},$$

where

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}.$$

The critical value at the 5% significance level is the value of t_c for $n_A + n_B - 2$ degrees of freedom in Table 5.

If the alternative hypothesis is H_1: $\mu_A > \mu_B$, we reject H_0 at the 5% significance level if $t > t_c$.

If the alternative hypothesis is H_1: $\mu_A < \mu_B$, we reject H_0 at the 5% significance level if $t < -t_c$.

Table 5 5% critical values for one-tailed Student's t-test

Degrees of freedom	Critical value (t_c)	Degrees of freedom	Critical value (t_c)
1	6.314	21	1.721
2	2.920	22	1.717
3	2.353	23	1.714
4	2.132	24	1.711
5	2.015	25	1.708
6	1.943	26	1.706
7	1.895	27	1.703
8	1.860	28	1.701
9	1.833	29	1.699
10	1.812	30	1.697
11	1.796	31	1.696
12	1.782	32	1.694
13	1.771	33	1.692
14	1.761	34	1.691
15	1.753	35	1.690
16	1.746	36	1.688
17	1.740	37	1.687
18	1.734	38	1.686
19	1.729	39	1.685
20	1.725	40	1.684

The result of a one-sided *un*matched pair test? Despite this being a cricket scoreboard, it shows the result of a 2003 *rugby* World Cup match: Australia 142 Namibia 0.

Example 15 A one-sided one-sample t-test

The average number of customers that a café serves during lunchtime on a weekday is 80.3. To try to increase this number, it starts to advertise regularly in the local newspaper. In the 20 weekdays following the start of the adverts, the average number of customers was 84.4, and the standard deviation of the number of customers was 9.2. The manager of the café wants to test whether advertising has changed the average number of lunchtime customers, or whether the difference between 80.3 and 84.4 is simply random variation. We have:

$$A = 80.3, \quad n = 20, \quad \overline{x} = 84.4, \quad s = 9.2.$$

The null hypothesis is

$$H_0 : \ \mu = 80.3,$$

and the test statistic is

$$t = \frac{\overline{x} - 80.3}{\text{ESE}} = \frac{84.4 - 80.3}{9.2/\sqrt{20}} \simeq 1.99.$$

The analysis up to this point is the same for a one-sided test as for a two-sided test. For a two-sided test we would specify the alternative hypothesis as H_0: $\mu \neq 80.3$. Suppose, though, that the manager was certain, even before gathering data, that newspaper advertising could do no harm – it would either increase custom or result in no change. Then the manager might choose to use a one-sided test and use the alternative hypothesis

$$H_1 : \mu > 80.3.$$

The number of degrees of freedom equals $n - 1 = 19$. From Table 5, the 5% critical value for a one-sided t-test with 19 degrees of freedom is 1.729. This is less than 1.99, the value of the test statistic. Also, the difference between $\overline{x}$ (84.4) and A (80.3) is in the direction consistent with H_1. Hence the null hypothesis is rejected at the 5% significance level.

The manager can conclude that there is moderate evidence that the average number of lunchtime customers has increased since advertising started.

In this next activity you are asked to perform a one-sided matched pairs t-test.

Activity 26 Benefit of exercise

An exercise physiologist measured the resting heart rate, in beats per minute, of seven people immediately before they started a one-year exercise program. These readings and their resting heart rate readings at the end of the program are given in Table 6. Assuming that the exercise program will not increase resting heart rate on average, examine the evidence that the exercise program reduces resting heart rate.

Table 6 Resting heart rates before and after the exercise program

	Resting heart rate	
Person	Before	After
1	74	71
2	71	68
3	68	66
4	75	72
5	75	70
6	72	73
7	69	67

Comparison of Tables 2 (Subsection 3.3) and 5 shows that the critical value for the one-sided test is less than for the two-sided test. This is true for every possible value of the degrees-of-freedom parameter. For example, when there are 20 degrees of freedom, the critical value is 2.086 for the two-sided test at the 5% significance level but only 1.725 for the one-sided test, and for 5 degrees of freedom it is 2.571 for the two-sided test but only 2.015 for the one-sided test.

Since one-sided tests set a lower threshold than two-sided tests, they are more likely to lead to the null hypothesis being rejected. You might have thought that this would make them popular, as experimenters typically hope to reject the null hypothesis. However, because they yield a lower threshold, an experimenter choosing to use a one-tailed test must be able to defend that choice against the question, *Why didn't you use a two-tailed test?* You should use a one-sided hypothesis test only when it is clear – before looking at the data – that if H_0 is wrong, then there is only one direction in which it can be wrong.

You have now covered the material related to Screencast 6 for Unit 10 (see the M140 website).

Exercise on Section 6

Exercise 9 Involving parents

The head teacher of a small primary school is keen to help the children to read more quickly and she hopes this can be achieved by involving parents. At the beginning of the school year she calls a meeting of the parents of the 23 children who have just started school and explains how they can help by listening to their children read at home.

The school considers that a child can read after he or she has successfully completed the first four books in the series they use to teach reading. That year, the number of days taken to learn to read had a mean of 119.2 days and a standard deviation of 29.6 days. In previous years, the number of days till a child could read had a mean of 127.3 days. The head teacher wishes to test whether involving parents affects the mean time in which a child learns to read.

(a) Give the null hypothesis, suggest why a one-sided alternative hypothesis could be considered appropriate, and give such an alternative hypothesis.

(b) Give the values of μ, n, $\overline{x}$ and s.

(c) Calculate the value of the test statistic.

(d) Say whether you reject the null hypothesis. What do you conclude?

7 Computer work: experiments

In this section, you will use Minitab to perform one-sample t-tests, two-sample t-tests and paired t-tests. You will also learn how to use Minitab to calculate confidence intervals that correspond to t-tests. You should now turn to the Computer Book and work through Chapter 10.

Summary

In this unit, we have briefly considered the nature of experiments and their role in advancing knowledge; different kinds of experiment were distinguished. You now (hopefully) know how to grow mustard seeds! You will have found that measuring root lengths is not easy, especially as they are not naturally straight. So you will have seen that measurements cannot be made with perfect accuracy; simply because a number is recorded to the nearest millimetre does not mean that the measurement is made with that level of accuracy.

You also met the family of t distributions and the degrees-of-freedom parameter that relates to them. You have learned a new hypothesis test – the two-sample t-test – and used it to analyse the data from your mustard seed experiment. The test requires an assumption that two populations have the same variance. You have used a rule of thumb for checking this assumption and a method of forming a pooled estimate of their common variance, if their variances can be assumed equal.

The t distribution was also used to perform one-sample tests when the sample size is too small for a z-test. This t-test extends easily to test hypotheses when data are from matched pairs, and you used it for that purpose.

In addition, you have learned to use critical values from t distributions to form confidence intervals for the mean of one population or for the difference between the means of two populations. In the latter case the data might be in the form of matched pairs or it might come from two unrelated samples. You also learned how to use t-tests with one-sided alternative hypotheses, although in most situations two-sided alternative hypotheses should be used.

Finally, you have also used Minitab to perform t-tests and construct confidence intervals based on t distributions.

Learning outcomes

After working through this unit, you should be able to:

- distinguish between experimental and non-experimental forms of inquiry
- distinguish between three kinds of experiments (exploratory, measurement and hypothesis testing)
- appreciate the requirements in setting up, maintaining and completing a small scientific experiment
- recognise both samples and areas of investigation for which it would be unwise to use the t-test
- examine whether it is reasonable to assume that two population variances are equal
- carry out a two-sample t-test for unrelated samples
- carry out a one-sample t-test
- carry out a matched-pairs t-test
- calculate confidence intervals for one-sample and two-sample data from populations satisfying the distributional conditions required by the t-tests
- test a one-sided alternative hypothesis
- use Minitab to perform t-tests and construct confidence intervals when sample sizes are small.

Solutions to activities

Solution to Activity 1

The chef could prepare two pies, one with and one without the new ingredient. Then a reasonable experiment is for a large number of people to try both pies. Each person says which pie they prefer, and the sign test (Unit 6) can be used to test whether one pie is better than the other. (Half the people should try the old pie first, and half should try the new pie first, in case the order in which they are tried matters.)

Solution to Activity 2

One possibility is to use two groups of people, as similar as possible with respect to those features which you think might be relevant to poetry appreciation, such as age, sex and educational background. Also, the people should not have read the poem before. Each person in one group would be asked to read the three-verse version of the poem and rate how good it is on a seven-point scale. Each person in the other group would be asked to read, and rate, the truncated two-verse version. The two batches of ratings could then be analysed.

Solution to Activity 3

(a) *Null hypothesis:* Microbes do not make food putrefy.

Alternative hypothesis: Microbes do make food putrefy.

Test: Prevent microbes from acting on the food.

Possible results and conclusions: The food does not putrefy; therefore reject the null hypothesis. The food putrefies; therefore the null hypothesis is supported.

(b) *Null hypothesis:* Sound is not transmitted by the jostling of air molecules.

Alternative hypothesis: Sound is transmitted by the jostling of air molecules.

Test: Ring a bell inside a vessel with all the air pumped out.

Possible results and conclusions: The bell is inaudible; therefore reject the null hypothesis. The bell can be heard; therefore the null hypothesis is supported.

Solution to Activity 4

(a) This activity is designed to answer specific questions by pursuing a specific investigation. Thus it is a scientific experiment if it is conducted properly. (It must use a method that is repeatable.)

(b) For the same reasons as the solution to (a), this is a scientific experiment – if it is conducted properly.

(c) This is not a scientific experiment, as no experiment is being conducted. An example of a scientific experiment would be to buy a particular brand of tea and compare it with your usual brand.

(d) For the same reasons as the solution to (a), this is a scientific experiment – if it is conducted properly.

Solution to Activity 5

In Activity 4, you should have identified three scientific experiments.

- Measuring the distance between the Earth and the Sun, which is a measurement experiment.
- Leaving work an hour later to see if it makes much difference to your travel time to get home, which is an exploratory experiment. It could also be rephrased as a hypothesis-testing experiment, as follows.

  Hypothesis: Leaving work an hour later makes a big difference to your average travel time to get home.

  Prediction: If you leave work an hour later, your average time to get home will change substantially.

  Test: Leave work an hour later for a month and see if your average journey time changes a lot.

  Possible results and conclusions: Average journey time almost unchanged; therefore hypothesis is false. Average journey time goes up (or down) by a lot; therefore hypothesis is supported.
- Investigating whether obesity is caused by overeating, which is a hypothesis-testing experiment. It seeks to test a hypothesis about the cause of a phenomenon, as follows.

  Hypothesis: Obesity is caused by overeating.

  Prediction: Eating a lot will cause people to be obese.

  Test: Measure the weights of people who eat a lot and people who do not, and compare these with their ideal weights, as defined by the BMI (body mass index), for example.

  Possible results and conclusions: People who eat a lot are no more obese than people who do not; therefore hypothesis is false. People who eat a lot are more obese than people who do not; therefore hypothesis is supported.

Solution to Activity 6

In Activity 5, you should have identified two hypothesis-testing experiments.

- Leaving work an hour later to see if it makes much difference to your travel time to get home. Treating this as a hypothesis-testing experiment gives the following.

  Null hypothesis: Leaving work an hour later makes no difference to your average travel time to get home.

  Alternative hypothesis: Leaving work an hour later makes a big difference to your average travel time to get home.

  Test: Leave work an hour later for a month and see if your average journey time changes a lot.

  Possible results and conclusions: Average journey time goes up (or down) by a lot; therefore reject the null hypothesis. Average journey time almost unchanged, so do not reject the null hypothesis.
- Investigating whether obesity is caused by overeating. The statistical hypothesis test is as follows.

  Null hypothesis: Obesity is not caused by overeating.

  Alternative hypothesis: Obesity is caused by overeating.

Test: Measure the weights of people who eat a lot and people who do not and compare these with their ideal weights, as defined by the BMI.

Possible results and conclusions: Clear evidence that people who eat a lot are more obese than people who do not; therefore reject the null hypothesis. Find that people who eat a lot are no more obese than people who do not; therefore the null hypothesis is not rejected.

Solution to Activity 7

The seedlings need to grow under conditions that are as similar as possible except for the presence or absence of light. It is especially important that the seedlings in the two groups are grown at the same temperature and humidity, and that they are equally spaced from each other, so that they do not suffer from unequal crowding effects. By growing the two sets of seedlings side by side in adjacent pots, it should be possible to control for temperature. By carefully spacing the seeds in the pots, it should be possible to avoid unequal crowding effects.

Darkness will be achieved for one set of seedlings by covering the pot in which they are growing with aluminium kitchen foil. As well as cutting out the light, this is likely to have the effect of raising the humidity in the enclosed air space. It is important that the seedlings grown in the light should experience similar humidity, and this can be achieved by covering their pot as well. The simplest solution is to cover the pot with a piece of clear plastic or a plastic bag. However, aluminium might conduct heat differently from clear plastic, so this might introduce an unwanted difference into the experiment. The difference in temperature conditions induced by the two different coverings will probably be small, especially if the room temperature is fairly constant during the experiment. It is up to you whether you try to devise ingenious ways of reducing this source of error, but it is not worth spending too long looking for ways of improving this particular aspect of the experiment.

There is another, quite subtle, factor that needs to be controlled. Light will affect the stems and leaves of the seedlings, and the stems and leaves of the two groups of seedlings will differ quite strikingly in their appearance as a result. It is possible that these changes in the leaves and stems themselves affect the roots. Perhaps a big stem stimulates the growth of a big root, whereas a small stem does not. If this is so, then you cannot be certain whether any differences which emerge are directly due to the effect of light on the roots, or whether they are due to the effect of the light on the leaves and stems, which in turn affect the growth of the roots. To control for this you should cut off the growing shoots from the seedlings. We suggest that you do not treat all the seedlings in this way. This will allow you to see how long the roots grow on the seedlings which are not cut (details in Subsection 2.5).

Solution to Activity 8

Null hypothesis Light has no effect on root growth.

Prediction based on null hypothesis The seedlings grown in the light will not differ in average root length from the seedlings grown in the dark.

Possible results and conclusions If there is a difference in average root lengths between the two groups of seedlings, then the null hypothesis must be rejected, and the alternative hypothesis – that light does influence root growth – should be accepted. If there is no difference, then the null hypothesis is supported.

Solution to Activity 9

We have two independent sample sets – one set grown in the light and the other in the dark. You have come across two tests that can be applied to two unrelated samples of data. The χ^2 test could be applied, if you were to construct a contingency table from the data by dividing the measurements into categories. However, this test needs fairly large samples, and unless almost all your seeds germinate, it is likely that your samples will not be large enough.

The two-sample z-test can also be used on unrelated samples of data, but, like the χ^2 test, it requires samples larger than those that you have. Thus a different test is needed.

Solution to Activity 10

(a) Sums, and sums of squares:

	x_A	x_A^2	x_B	x_B^2
	2	4	8	64
	7	49	11	121
	8	64	3	9
	3	9	5	25
	5	25	8	64
$\sum$	25	151	35	283

For Group A, $\overline{x}_A = 25/5 = 5$. Also

$$\sum x_A^2 - \frac{(\sum x_A)^2}{n_A} = 151 - \frac{25^2}{5} = 26,$$

so

$$s_A^2 = \frac{26}{n_A - 1} = \frac{26}{4} = 6.5.$$

For Group B, $\overline{x}_B = 35/5 = 7$. Also

$$\sum x_B^2 - \frac{(\sum x_B)^2}{n_B} = 283 - \frac{35^2}{5} = 38,$$

so

$$s_B^2 = \frac{38}{n_B - 1} = \frac{38}{4} = 9.5.$$

(b) The two sample variances are 6.5 and 9.5. Dividing the larger by the smaller gives $9.5/6.5 \simeq 1.462$. As this is less than 3, it is reasonable to accept that the populations have the same variance.

(c) The pooled estimate of the common variance is

$$s_{\mathsf{p}}^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$
$$= \frac{(5-1) \times 6.5 + (5-1) \times 9.5}{5 + 5 - 2} = \frac{64}{8} = 8.$$

Hence s_{p} (the pooled estimate of the common standard deviation) is $\sqrt{8} \simeq 2.8284$.

Solution to Activity 11

Yes. The value -3.906 calculated for the test statistic t is less than -2.571 (i.e. it is more extreme than the critical value), so the null hypothesis should be rejected at the 5% significance level. Thus there is moderate evidence that the average weight gain of calves differs between the two diets. As the mean weight gains were 0.5125 in Group A and 0.6767 in Group B, there is some evidence that average weight gain is higher on diet B.

Solution to Activity 12

From Activity 10, $\overline{x}_A = 5$, $\overline{x}_B = 7$ and $s_\mathsf{p} \simeq 2.8284$.

$$\begin{aligned} t &= \frac{\overline{x}_A - \overline{x}_B}{s_\mathsf{p}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} \\ &\simeq \frac{5-7}{2.8284\sqrt{\dfrac{1}{5} + \dfrac{1}{5}}} \\ &\simeq \frac{-2}{1.7889} \\ &= -1.118 \quad \text{(rounded to three decimal places).} \end{aligned}$$

The number of degrees of freedom is $5 + 5 - 2 = 8$. From Table 2 (Subsection 3.3) the critical value is $t_\mathsf{c} = 2.306$. So we reject H_0 in favour of H_1 if $t \geq 2.306$ or $t \leq -2.306$. The value of the test statistic -1.118 is nearer to zero than the critical value 2.306, so we cannot reject the null hypothesis (of no difference between the population means) at the 5% significance level. Hence there is little evidence that the time taken to manoeuvre the ball around the obstacle course is influenced by whether the child saw the course in advance or was told in advance how to negotiate it.

Solution to Activity 13

The null hypothesis is

$H_0 : \mu_A = \mu_B$

(the population mean heights of the two varieties are equal),

against the alternative hypothesis

$H_1 : \mu_A \neq \mu_B$

(the population mean heights of the two varieties are not equal).

First we check that it is reasonable to assume a common population variance. The two sample variances are $s_A^2 = 0.051^2 = 0.002\,601$ and $s_B^2 = 0.038^2 = 0.001\,444$. Dividing the larger sample variance by the smaller sample variance gives $0.002\,601/0.001\,444 \simeq 1.801$. This is much less than 3, so assuming a common population variance seems reasonable.

It is estimated as:

$$\begin{aligned} s_\mathsf{p}^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \\ &= \frac{4 \times 0.002\,601 + 5 \times 0.001\,444}{5 + 6 - 2} \\ &= \frac{0.017\,624}{9} \\ &\simeq 0.001\,958\,2. \end{aligned}$$

(This is between 0.001 444 and 0.002 601, so there is no obvious calculation error.) We have $s_\mathsf{p} \simeq \sqrt{0.001\,958\,2} \simeq 0.044\,252$.

Now

$$\begin{aligned} t &= \frac{\overline{x}_A - \overline{x}_B}{s_\mathrm{p}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} \\ &\simeq \frac{1.252 - 1.023}{0.044\,252\sqrt{\dfrac{1}{5} + \dfrac{1}{6}}} \\ &\simeq \frac{0.229}{0.026\,796} \\ &\simeq 8.546. \end{aligned}$$

The number of degrees of freedom is 9, so from Table 2 (Subsection 3.3), the critical value $t_\mathrm{c} = 2.262$. Hence H_0 is easily rejected at the 5% significance level. There is evidence that the varieties differ in their average heights.

Solution to Activity 14

Stemplot for the *light* sample (Group A):

	1	3 7
	2	1 9
	3	1 9 9
	4	
	5	0 2 5
$n = 10$	1	3 represents 13 mm

Stemplot for the *dark* sample (Group B):

	1	4 6 7
	2	0 2 2 8
	3	2 6
	4	1
$n = 10$	1	4 represents 14 mm

Solution to Activity 15

For the sample grown in the light (Group A),

$$\sum x_A = 21 + 39 + 31 + \cdots + 17 = 346$$

and

$$\sum x_A^2 = 21^2 + 39^2 + 31^2 + \cdots + 17^2 = 13\,972.$$

As $n_A = 10$,

$$\sum x_A^2 - \frac{(\sum x_A)^2}{n_A} = 13\,972 - \frac{346^2}{10} = 2000.4,$$

so

$$s_A^2 = \frac{2000.4}{n_A - 1} = \frac{2000.4}{9} \simeq 222.267.$$

For the sample grown in the dark (Group B),

$$\sum x_B = 22 + 16 + 20 + \cdots + 22 = 248$$

and

$$\sum x_B^2 = 22^2 + 16^2 + 20^2 + \cdots + 22^2 = 6894.$$

As $n_B = 10$ as well,

$$\sum x_B^2 - \frac{(\sum x_B)^2}{n_B} = 6894 - \frac{248^2}{10} = 743.6,$$

so

$$s_B^2 = \frac{743.6}{n_B - 1} = \frac{743.6}{9} \simeq 82.622.$$

Dividing the larger sample variance by the smaller sample variance gives $222.267/82.622 \simeq 2.69$. As this is less than 3, our rule of thumb says we can assume the samples are from populations whose variances are equal.

Solution to Activity 16

The sample size is $n = 5$. We have

$$\sum x = 3.6 + 3.2 + 3.1 + 2.6 + 3.9 = 16.4,$$

so the sample mean is

$$\overline{x} = \frac{\sum x}{n} = \frac{16.4}{5} = 3.28.$$

Also,

$$\sum x^2 = 3.6^2 + 3.2^2 + 3.1^2 + 2.6^2 + 3.9^2 = 54.78,$$

so

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) = \frac{1}{5-1}\left(54.78 - \frac{16.4^2}{5}\right) \\ &= \frac{1}{4}(54.78 - 53.792) = 0.247. \end{aligned}$$

Thus, the standard deviation is $s = \sqrt{0.247} \simeq 0.496\,99$.

Solution to Activity 17

(a) The test statistic for a one-sample z-test is

$$z = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

For the tomato grower's experiment,

$$\text{ESE} \simeq \frac{0.496\,99}{\sqrt{5}} \simeq 0.222\,26, \quad \text{so } z \simeq \frac{3.28 - 4}{0.222\,26} \simeq -3.239.$$

(b) The z-test is not appropriate here for the same reason that it was not appropriate in Section 3. The sample size is so small that $s/\sqrt{n}$ may not be a good estimate of the standard error. Consequently, for such a small sample size, this test statistic has not got a standard normal distribution.

Solution to Activity 18

The sample size n is 5, so the number of degrees of freedom is 4. Thus from Table 2 (Subsection 3.3) the critical value is 2.776.

Solution to Activity 19

(a) Some patients might be more responsive to treatment than other patients. To reduce the effect of variation between patients, each patient received *both* drugs.

(b) The order in which a patient receives treatments might have an effect – perhaps patients tend to respond more to the first treatment they receive. To reduce this effect, each of the two drugs was the first that five patients received and the second that the other five patients received.

Solution to Activity 20

Patient	Form L	Form R	Difference $L - R$
1	+1.9	+0.7	+1.2
2	+0.8	-1.6	+2.4
3	+1.1	-0.2	+1.3
4	+0.1	-1.2	+1.3
5	-0.1	-0.1	0.0
6	+4.4	+3.4	+1.0
7	+5.5	+3.7	+1.8
8	+1.6	+0.8	+0.8
9	+4.6	0.0	+4.6
10	+3.4	+2.0	+1.4

Solution to Activity 21

The hypotheses are:

$$H_0 : \mu_d = 0,$$
$$H_1 : \mu_d \neq 0.$$

Let d denote the difference $L - R$. First it is necessary to calculate the mean ($\overline{d}$) and standard deviation (s) for the sample of differences.

	d	d^2
	1.2	1.44
	2.4	5.76
	1.3	1.69
	1.3	1.69
	0.0	0.00
	1.0	1.00
	1.8	3.24
	0.8	0.64
	4.6	21.16
	1.4	1.96
$\sum$	15.8	38.58

Thus, writing d instead of x in the formulas, the mean of these differences is

$$\overline{d} = \frac{\sum d}{n} = \frac{15.8}{10} = 1.58$$

and the variance is

$$s^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right) = \frac{1}{10-1}\left(38.58 - \frac{15.8^2}{10}\right)$$
$$= \frac{1}{9}(38.58 - 24.964) \simeq 1.512\,89.$$

Thus, the standard deviation is $s \simeq \sqrt{1.512\,89} \simeq 1.230\,00$. Hence,

$$\text{ESE} = \frac{s}{\sqrt{n}} \simeq \frac{1.230\,00}{\sqrt{10}} \simeq 0.388\,96.$$

Now, the test statistic is

$$t = \frac{\overline{d} - A}{\text{ESE}}$$

but $A = 0$ since the null hypothesis is H_0: $\mu_d = 0$. Thus

$$t = \frac{1.58 - 0}{0.388\,96} \simeq 4.062.$$

The critical value is obtained from Table 2 (Subsection 3.3). The number of degrees of freedom is $n - 1 = 10 - 1 = 9$, so the critical value is 2.262. The value of t is 4.062, which is greater than 2.262. So H_0 is rejected at the 5% significance level.

The conclusion is that the population mean difference is not zero. Thus, on the basis of this experiment it seems that the two forms of hyoscyamine hydrobromide do differ in their effectiveness at increasing sleep. Because the test statistic is positive, there is moderate evidence that form L gives a greater gain in sleep on average.

Solution to Activity 22

	Weight in grams		Difference	
Object	Equipment A	Equipment B	d	d^2
1	3.6	3.3	0.3	0.09
2	4.3	4.4	-0.1	0.01
3	11.4	11.2	0.2	0.04
4	15.9	15.5	0.4	0.16
5	16.4	16.6	-0.2	0.04
6	18.7	18.7	0.0	0.00
7	21.1	20.7	0.4	0.16
8	21.8	21.4	0.4	0.16
9	24.1	23.8	0.3	0.09
$\sum$			1.7	0.75

(a) The mean of the differences is

$$\overline{d} = \frac{\sum d}{n} = \frac{1.7}{9} \simeq 0.188\,89.$$

Also

$$s^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right) = \frac{1}{9-1}\left(0.75 - \frac{1.7^2}{9}\right)$$
$$\simeq \frac{1}{8}(0.75 - 0.321\,11) \simeq 0.053\,611.$$

Thus, the standard deviation is $s \simeq \sqrt{0.053\,611} \simeq 0.231\,54$.

(b) The hypotheses are

$$H_0 : \mu_d = 0,$$
$$H_1 : \mu_d \neq 0.$$

(c) The estimated standard error is

$$\text{ESE} = \frac{s}{\sqrt{n}} \simeq \frac{0.231\,54}{\sqrt{9}} = 0.077\,18,$$

so the test statistic is

$$t = \frac{\overline{d}}{\text{ESE}} \simeq \frac{0.188\,89}{0.077\,18} \simeq 2.447.$$

(d) There are $n - 1 = 8$ degrees of freedom. From Table 2 (Subsection 3.3), the 5% critical value (t_c) for 8 degrees of freedom is 2.306. The test statistic, 2.447, is greater than 2.306, so H_0 is rejected at the 5% significance level.

(e) There is moderate evidence that the population mean difference is not zero. This suggests that the two pieces of equipment systematically differ in the weights they give – A seems, on average, to give higher weights than B, which means that the weighing machines cannot both be unbiased. There is moderate evidence that either equipment A on average gives weights that are too high, or that B on average gives weights that are too low, or that both pieces of equipment are biased (when we do not know the direction or directions of bias).

Solution to Activity 23

The number of degrees of freedom is $n - 1 = 15 - 1 = 14$. From Table 2 (Subsection 3.3), the critical value for 14 degrees of freedom is 2.145.

From this we calculate

$$t_c \frac{s}{\sqrt{n}} = 2.145 \times \frac{1.18}{\sqrt{15}} \simeq 0.65.$$

Thus a 95% confidence interval for the height of sunflowers, in metres, is $(3.60 - 0.65, 3.60 + 0.65)$. This is $(2.95, 4.25)$.

Solution to Activity 24

From the solution to Activity 21, $n = 10$, $\overline{d} = 1.58$ and $\text{ESE} = s/\sqrt{n} \simeq 1.230\,00/\sqrt{10} \simeq 0.388\,96$. Also, there are $n - 1 = 9$ degrees of freedom, for which the critical value is $t_c = 2.262$.

From this we calculate

$$t_c \frac{s}{\sqrt{n}} = t_c \times \text{ESE} \simeq 2.262 \times 0.388\,96 \simeq 0.88.$$

Thus a 95% confidence interval for the mean number of hours sleep gained is $(1.58 - 0.88, 1.58 + 0.88)$. This is $(0.70, 2.46)$.

Solution to Activity 25

We first calculate the two sample variances:

$$s_A^2 = 1.18^2 = 1.3924 \quad \text{and} \quad s_B^2 = 1.09^2 = 1.1881.$$

Dividing the larger sample variance by the smaller sample variance gives $1.3924/1.1881 \simeq 1.172$. This is much less than 3, so from our rule of thumb it is reasonable to assume a common population variance.

We estimate this common variance by pooling the sample variances:

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{(15 - 1) \times 1.3924 + (20 - 1) \times 1.1881}{15 + 20 - 2}$$
$$= \frac{42.0675}{33} \simeq 1.2748.$$

So $s_p \simeq 1.1291$.

The number of degrees of freedom is $n_A + n_B - 2 = 15 + 20 - 2 = 33$. From Table 2 (Subsection 3.3) the critical value for 33 degrees of freedom is 2.035.

We first calculate

$$t_c s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \simeq 2.035 \times 1.1291 \sqrt{\frac{1}{15} + \frac{1}{20}} \simeq 0.78$$

and

$$\overline{x}_A - \overline{x}_B = 3.60 - 2.76 = 0.84.$$

Thus the confidence interval for the mean difference in height between the heights of sunflowers, in metres, is $(0.84 - 0.78, 0.84 + 0.78)$, which is $(0.06, 1.62)$.

Solution to Activity 26

Let μ_d denote the mean change in resting heart rate over the course of the exercise program. The null hypothesis is that there is no change in average resting heart rate:

$$H_0 : \mu_d = 0.$$

As it is assumed that the exercise program will not increase resting heart rate, the alternative hypothesis is one-sided:

$$H_1 : \mu_d < 0.$$

The data are paired (there are two values for each person), so a matched-pairs t-test is appropriate. The differences d (after $-$ before) are calculated, and then the sample mean and sample variance of d.

Person	Resting heart rate Before	After	Difference d	d^2
1	74	71	-3	9
2	71	68	-3	9
3	68	66	-2	4
4	75	72	-3	9
5	75	70	-5	25
6	72	73	1	1
7	69	67	-2	4
$\sum$			-17	61

The mean of the differences is

$$\overline{d} = \frac{\sum d}{n} = \frac{-17}{7} \simeq -2.4286.$$

Also

$$s^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right) = \frac{1}{7-1}\left(61 - \frac{(-17)^2}{7}\right)$$
$$\simeq \frac{1}{6}(61 - 41.285\,71) \simeq 3.285\,71.$$

Thus, the standard deviation is $s \simeq \sqrt{3.285\,71} \simeq 1.812\,65$.

The estimated standard error is

$$\text{ESE} = \frac{s}{\sqrt{n}} \simeq \frac{1.812\,65}{\sqrt{7}} \simeq 0.685\,12,$$

so the test statistic is

$$t = \frac{\overline{d}}{\text{ESE}} \simeq \frac{-2.4286}{0.685\,12} \simeq -3.545.$$

There are $n - 1 = 6$ degrees of freedom. From Table 5 (Section 6), the 5% critical value (t_c) for six degrees of freedom is 1.943. As -3.545 is less than -1.943, H_0 is rejected at the 5% significance level. Thus there is moderate evidence that the exercise program is associated with a reduction in resting heart rate. (Of course, it may be that the people taking the exercise program also started to have a healthier lifestyle in other ways too, such as a change of diet.)

Solutions to exercises

Solution to Exercise 1

(a) This is designed to answer a specific question by pursuing a specific investigation. Thus it is a scientific experiment provided that it is properly conducted.

(b) This is not a scientific experiment: there is no experiment taking place. It is about gathering information rather than exploring, measuring or testing a hypothesis.

Solution to Exercise 2

(a) Driving a car with all the windows open to see whether petrol consumption is affected is an exploratory experiment. It can be rephrased as a hypothesis-testing experiment, as follows.

Hypothesis: Driving with all the windows open affects petrol consumption.

Prediction: If I drive with the windows open, petrol consumption will change (assuming the windows are usually closed).

Test: See what happens to petrol consumption when you drive with the windows open.

Possible results and conclusions: Petrol consumption almost unchanged, so hypothesis is false. Petrol consumption changes substantially, so hypothesis supported.

(b) Formulating this as a statistical hypothesis test gives the following.

Null hypothesis: Driving with all the windows open does not affect petrol consumption.

Alternative hypothesis: Driving with all the windows open affects petrol consumption.

Test: See what happens to petrol consumption when driving with the windows open.

Possible results and conclusions: Petrol consumption changes substantially; therefore reject the null hypothesis. Petrol consumption almost unchanged, so do not reject the null hypothesis.

Solution to Exercise 3

The null hypothesis is

$$H_0 : \mu_H = \mu_B \quad \text{(average weight gain is the same on the two diets)},$$

against the alternative hypothesis

$$H_1 : \mu_H \neq \mu_B \quad \text{(average weight gain differs between the two diets)}.$$

First we check that it is reasonable to assume a common population variance. The two sample variances are $s_H^2 = 0.081^2 = 0.006\,561$ and $s_B^2 = 0.088^2 = 0.007\,744$. Dividing the larger sample variance by the smaller sample variance gives $0.007\,744/0.006\,561 = 1.180$ (rounded to three decimal places). This is much less than 3, so assuming a common population variance is reasonable.

It is estimated as:

$$\begin{aligned} s_{\mathsf{p}}^2 &= \frac{(n_H-1)s_H^2+(n_B-1)s_B^2}{n_H+n_B-2} \\ &= \frac{19\times 0.006\,561+18\times 0.007\,744}{20+19-2} \\ &= \frac{0.264\,051}{37} \\ &\simeq 0.007\,136\,5. \end{aligned}$$

(This is between 0.006 561 and 0.007 744, so there is no obvious calculation error.) We have $s_{\mathsf{p}} \simeq \sqrt{0.007\,136\,5} \simeq 0.084\,478$.

Hence

$$\begin{aligned} t &= \frac{\overline{x}_H-\overline{x}_B}{s_{\mathsf{p}}\sqrt{\dfrac{1}{n_H}+\dfrac{1}{n_B}}} \\ &\simeq \frac{0.542-0.554}{0.084\,478\sqrt{\dfrac{1}{20}+\dfrac{1}{19}}} \\ &\simeq \frac{-0.012}{0.027\,063\,5} \\ &\simeq -0.443. \end{aligned}$$

The number of degrees of freedom is $20+19-2=37$, so from Table 2 (Subsection 3.3), the critical value $t_{\mathsf{c}}=2.026$. As -0.443 is (much) closer than the critical value to 0, H_0 is not rejected at the 5% significance level. There is little evidence that average weight gain differs between the two diets.

Solution to Exercise 4

The null hypothesis is

$$H_0 : \mu_A = \mu_B \quad \text{(average weight is the same in the two production lines)},$$

against the alternative hypothesis

$$H_1 : \mu_A \neq \mu_B \quad \text{(average weight differs between the two production lines)}.$$

First we check that it is reasonable to assume a common population variance. The two sample variances are $s_A^2 = 3.58^2 = 12.8164$ and $s_B^2 = 4.73^2 = 22.3729$. Dividing the larger sample variance by the smaller sample variance gives $22.3729/12.8164 = 1.746$ (rounded to three decimal places). This is less than 3, so we will assume a common population variance.

The population variance is estimated as:

$$\begin{aligned} s_{\mathsf{p}}^2 &= \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2} \\ &= \frac{14\times 12.8164+14\times 22.3729}{15+15-2} \\ &= \frac{492.6502}{28} \\ &\simeq 17.595. \end{aligned}$$

(This is between 12.8164 and 22.3729, as it should be.) We have $s_{\mathsf{p}} \simeq \sqrt{17.595} \simeq 4.1946$.

Hence

$$\begin{aligned} t &= \frac{\overline{x}_A - \overline{x}_B}{s_{\mathrm{p}}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \\ &\simeq \frac{309.8 - 305.2}{4.1946\sqrt{\frac{1}{15} + \frac{1}{15}}} \\ &\simeq \frac{4.6}{1.532} \\ &\simeq 3.00. \end{aligned}$$

The number of degrees of freedom is $15 + 15 - 2 = 28$. Thus, from Table 2 (Subsection 3.3), the critical value $t_{\mathrm{c}} = 2.048$. As 3.00 is greater than 2.048, H_0 is rejected at the 5% significance level. There is moderate evidence that average weight differs between the two production lines – production line A seems to give a higher average weight than production line B.

Solution to Exercise 5

(a) The sample standard deviation, s, is $\sqrt{3.7} \simeq 1.9235$, so

$$\mathrm{ESE} = \frac{s}{\sqrt{n}} \simeq \frac{1.9235}{\sqrt{12}} \simeq 0.555\,27.$$

Hence the test statistic for the t-test is

$$t = \frac{\overline{x} - A}{\mathrm{ESE}} \simeq \frac{8.2 - 10}{0.555\,27} \simeq -3.242.$$

The number of degrees of freedom is $n - 1 = 11$. Thus from Table 2 (Subsection 3.3) the critical value for the 5% significance level is 2.201. As -3.242 is less than -2.201, the null hypothesis is rejected at the 5% significance level. There is evidence that the population mean does not equal 10.

(b) From part (a), we have that ESE $\simeq 0.555\,27$. When $A = 7.5$,

$$t = \frac{\overline{x} - A}{\mathrm{ESE}} \simeq \frac{8.2 - 7.5}{0.555\,27} \simeq 1.261.$$

There are still 11 degrees of freedom, so the critical value is still 2.201. As 1.261 is nearer than 2.201 to 0, the null hypothesis is not rejected at the 5% significance level. There is little evidence against the hypothesis that the population mean is 7.5.

Solution to Exercise 6

The relevant null and alternative hypotheses are:

$$H_0 : \mu_{N-O} = 0$$
$$H_1 : \mu_{N-O} \neq 0.$$

We need to analyse the sample of five differences $(N - O)$:

	d	d^2
	-5	25
	-8	64
	0	0
	-2	4
	+3	9
$\sum$	-12	102

Thus the mean difference $\overline{d} = \dfrac{\sum d}{n} = \dfrac{-12}{5} = -2.4$.

Thus

$$s^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right) = \frac{1}{5-1}\left(102 - \frac{(-12)^2}{5}\right)$$
$$= \frac{1}{4}(102 - 28.8) = 18.3,$$

and the standard deviation is $s = \sqrt{18.3} \simeq 4.2778$. Hence

$$\text{ESE} = \frac{s}{\sqrt{n}} = \frac{4.2778}{\sqrt{5}} \simeq 1.913\,11,$$

and

$$t = \frac{\overline{d}}{\text{ESE}} \simeq \frac{-2.4}{1.913\,11} \simeq -1.255.$$

The number of degrees of freedom is $n - 1 = 4$. Thus the critical value is 2.776. The test statistic -1.255 is closer to zero than the critical value 2.776, so we cannot reject the null hypothesis. Thus there is little evidence of a difference between the new drug and the old drug in their effect on the weight of male patients.

Solution to Exercise 7

We have a matched-pairs sample of data.

From the solution to Activity 22, $n = 9$, $\overline{x} \simeq 0.188\,89$ and $\text{ESE} = s/\sqrt{n} \simeq 0.231\,54/\sqrt{9} = 0.077\,18$.

Also, there are $n - 1 = 8$ degrees of freedom, for which the critical value is $t_c = 2.306$.

From this we calculate

$$t_c \frac{s}{\sqrt{n}} = t_c \times \text{ESE} \simeq 2.306 \times 0.077\,180 \simeq 0.18.$$

Thus a 95% confidence interval for the mean difference in weight, in grams, is $(0.19 - 0.18, 0.19 + 0.18)$. This is $(0.01, 0.37)$.

Solution to Exercise 8

We have two unrelated samples of data. In Exercise 4 we checked that it is reasonable to treat the population variances as being equal. From Exercise 4:

$$n_A = 15, \quad n_B = 15, \quad \overline{x}_A = 309.8, \quad \overline{x}_B = 305.2,$$
$$s_p \simeq 4.1946, \quad t_c = 2.048.$$

To apply the above formula, we first calculate

$$t_c s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \simeq 2.048 \times 4.1946 \sqrt{\frac{1}{15} + \frac{1}{15}} \simeq 3.1$$

and

$$\overline{x}_A - \overline{x}_B = 309.8 - 305.2 = 4.6.$$

Thus the confidence interval for the mean difference in weight of biscuits, in grams, is $(4.6 - 3.1, 4.6 + 3.1)$, which is $(1.5, 7.7)$.

Solution to Exercise 9

(a) The null hypothesis is that involving parents has not changed the mean time to learn to read, so that it is still 127.3 days. Thus

$$H_0 : \mu = 127.3.$$

If it is believed that involving parents could not hinder a child learning to read, then μ cannot be more than 127.3, giving the one-sided alternative hypothesis

$$H_1 : \mu < 127.3.$$

(b) $\mu = 127.3$ (under H_0), $n = 23$, $\overline{x} = 119.2$ and $s = 29.6$.

(c) $t = \dfrac{\overline{x} - \mu}{\text{SE}} = \dfrac{119.2 - 127.3}{29.6/\sqrt{23}} \simeq -1.312.$

(d) There are $n - 1 = 22$ degrees of freedom. As the test is one-sided, the critical value for the 5% significance level with 22 degrees of freedom is -1.717, from Table 5. The value $-1.312 > -1.717$, so we do not reject H_0 at the 5% significance level. Thus the experiment provides little evidence that involving parents reduces the average time taken by children to learn to read.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover image: Minxlj/www.flickr.com/photos/minxlj/422472167/. This file is licensed under the Creative Commons Attribution-Non commercial-No Derivatives Licence http://creativecommons.org/licenses/by-nc-nd/3.0/

Subsection 1.2 cartoon: www.funnytimes.com

Subsection 1.2 photo (crying baby): Microsoft Corporation

Subsection 1.2 photo (frozen fruit): Eskymaks / www.shutterstock.com

Subsection 1.2 figure (spraying against malaria), taken from: www.flickr.com/photos/field_museum_library/3608428064/sizes/o/in/photostream/

Subsection 1.3 photo (tea tasting): Janet Leigh / http://www.flickr.com/photos/eastleighnet/8952234371/sizes/o/in/photostream/

Subsection 1.3 photo (Karl Popper), taken from: http://ciudadves.blogspot.co.uk/2011/07/la-falsabilidad-de-karl-popper.html#!/2011/07/la-falsabilidad-de-karl-popper.html

Subsection 3.2 cartoon (null hypothesis): www.xkcd.com. This file is licensed under the Creative Commons Attribution-Noncommercial Licence http://creativecommons.org/licenses/by-nc/3.0/

Subsection 3.3 photo (tea tasting, t-test): Thunderbolt / www.flickr.com/photos/darjeelingteas/7221569558/sizes/l/in/photostream/. This file is licensed under the Creative Commons Attribution-Noncommercial-NoDerivatives Licence http://creativecommons.org/licenses/by-nc-nd/3.0/

Subsection 3.3 photo (*Lupinus hartwegii*): Miya.m / http://commons.wikimedia.org/ wiki/File:Lupinus_in_Hokkaido_20080630.jpg. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Exercises on Section 3, photo taken from: http://toronto.architectureforhumanity.org/events/2671

Subsection 4.2 cartoon (an unmatched pair): www.cartoonstock.com

Section 6 cartoon (multiple choice): www.cartoonstock.com

Section 6 photo (scoreboard): Hamish Blair / Getty Images

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index